DOCUMENT RESUME

ED 365 142 FL 021 737

AUTHOR	Stansfield, Charles W., Ed.
TITLE	Technology and Language Testing. A Collection of Papers from the Annual Colloquium on Language Testing Research (7th, Princeton, New Jersey, April 6-9, 1985).
INSTITUTION	Teachers of English to Speakers of Other Languages.
PUB DATE	86
NOTE	188p.
AVAILABLE FROM	TESOL, 1600 Cameron Street, Suite 300, Alexandria, VA 22314.
PUB TYPE	Books (010) -- Collected Works - General (020)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Chinese; Cloze Procedure; *Computer Assisted Testing; Computer Software; Difficulty Level; *English (Second Language); Heterogeneous Grouping; Higher Education; Interrater Reliability; Item Banks; Language Proficiency; *Language Tests; *Measurement Techniques; Program Descriptions; Reading Comprehension; Second Language Instruction; Second Language Learning; Speech Skills; Statistical Analysis; Testing; Testing Problems; Test Items; Trend Analysis; Writing Evaluation; Writing Instruction
IDENTIFIERS	ACTFL Proficiency Guidelines; Rasch Model; University of California Los Angeles

ABSTRACT

This collection of essays on measurement theory and language testing includes: "Computerized Adaptive Testing: Implications for Language Test Developers" (Peter Tung); "The Promise and Threat of Computerized Adaptive Assessment of Reading Comprehension" (Michael Canale); "Computerized Rasch Analysis of Item Bias in ESL Tests" (Harold S. Madsen, Jerry W. Larson); "Item Banking via DBASE II: The UCLA ESL Proficiency Examination Experience" (Grant Henning); "Computerized Multilevel ESL Testing, a Rapid Screening Methodology" (Marilyn M. Hicks); "Item Selection from Pretests in Mixed Ability Groups" (John H. A. L. deJong); "Correlation Between Human and Machine Ratings of Test of Spoken English Reading Passages" (Garry Molholt, Ari M. Presler); Development of a Tape-Mediated, ACTFL/ILR Scale-Based Test of Chinese Speaking Proficiency" (John L. D. Clark); "Using Technology to Assess Second Language Proficiency Through Cloze-Elide Tests" (Winton H. Manning); and "Using the Writer's Workbench in Composition Teaching and Testing" (Joy Reid). (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 365 142

Technology and Language Testing

Edited by
CHARLES W. STANSFIELD

FL021737

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

Kornblum

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TEACHERS OF ENGLISH TO SPEAKERS OF OTHER LANGUAGES

BEST COPY AVAILABLE

Technology and Language Testing

A collection of papers from the Seventh Annual Language Testing Research Colloquium, held at Educational Testing Service, Princeton, New Jersey, April 6-9, 1985

Edited by

Charles W. Stansfield

Teachers of English to Speakers of Other Languages

3

Copyright © 1986 by
the Teachers of English to Speakers of Other Languages
Washington, D.C.

All rights reserved. No part of this book may be copied or reproduced in any form or by any means without written permission of Teachers of English to Speakers of Other Languages.

Printed in the United States of America

TESOL Publications Manager, Julia Frank-McNeil
Composition by Wordworks, Lovettsville, Virginia
Printing by Pantagraph Printing, Bloomington, Illinois

ACKNOWLEDGEMENTS

Bringing to fruition a book of this nature invariably involves many people. It is appropriate to acknowledge the efforts of at least the major contributors here.

The volume is the product of a conference that was held at Educational Testing Service in April 1985. In coordinating this conference I received support or assistance from ETS staff members Robert Altman, Cathy Clauss, Vera Jones, and Russell Webster. In the process of editing the volume I was aided by Mare Fallon and Ruth Miller, and by the TESOL Publications Manager, Julia Frank-McNeil.

The volume could not have been published without the support of the TESOL Publications Committee. In particular, I would like to express my gratitude to Doug Brown, Chair of the Committee, for his helpful communication over the past two years.

Finally, I wish to thank the authors of these papers for working with me through various drafts and for granting me permission to include their work in this volume.

Contents

Introduction *Charles W. Stansfield* 3

Section 1. New developments in measurement theory: computerized adaptive testing and the application of latent trait models to test and item analysis 9

1. Computerized adaptive testing: implications for language test developers *Peter Tung* 11

2. The promise and threat of computerized adaptive assessment of reading comprehension *Michael Canale* 29

3. Computerized Rasch analysis of item bias in ESL tests *Harold S. Madsen and Jerry W. Larson* 47

4. Item banking via DBASE II: the UCLA ESL Proficiency Examination experience *Grant Henning* 69

5. Computerized multilevel ESL testing, a rapid screening methodology *Marilyn M. Hicks* 79

6. Item selection from pretests in mixed ability groups *John H. A. L. de Jong* 91

Section 2. Using technology to develop new measures of speaking, reading, and writing 109

7. Correlation between human and machine ratings of Test of Spoken English reading passages *Garry Molholt and Ari M. Presler* 111

8. Development of a tape-mediated, ACTFL/ILR scale-based test of Chinese speaking proficiency *John L. D. Clark* 129

9. Using technology to assess second language proficiency through cloze-elide tests. *Winton H. Manning* 147

10. Using the Writer's Workbench in composition teaching and testing *Joy Reid* 167

Technology and Language Testing

Introduction

For the past seven years, a group of applied linguists whose chief academic interest is the theory and practice of second language testing has met for two or more days either concurrent with, or immediately prior to, the annual TESOL Convention. The purpose of this meeting has been to present a collection of scholarly papers on language testing and to discuss issues of mutual interest. The meeting has come to be known as the Language Testing Research Colloquium. Many excellent papers have been presented at the colloquium. Three collections of them have been published previously (Briere & Hinofotis, 1979; Palmer, Groot & Trosper, 1981; Jones, Desbrisay & Paribakhp, 1985).

The group that attends the Colloquium is not formally constituted. The group has no office, no elected leadership, and does not publish a newsletter. Moreover, its members (if they can so be called) prefer its informal status and have consistently resisted the urge to convert this network of colleagues into an organization. Each year one or more members of the network volunteer to make arrangements for the next meeting, and to mail the new program to past participants. Usually, the individuals who volunteer to coordinate a meeting of the colloquium live near the site of the future meeting of TESOL. So in 1983, while attending the fifth colloquium held at the University of Ottawa, I was asked by several participants to make arrangements for the seventh colloquium, which it was hoped could be held at Educational Testing Service immediately prior to the Nineteenth Annual TESOL Convention in New York City.

Because of my work at Educational Testing Service, I was acutely aware of the importance to our field of new technological developments, both in measurement theory and in methods of test administration. Therefore, I suggested that the seventh colloquium focus on these developments and urged members to explore them in their papers.

When examining the history of educational measurement, one notes that until the beginning of the twentieth century, most tests were individually administered, in a face-to-face oral format, and in many cases the questions were tailored to the individual being examined. This format is still used at U.S. universities when a student takes comprehensive examinations as part of a process leading to a graduate degree. Outside of North America, individual examinations are often part of the process of applying for a position as a teacher or professor.

However, with the advent of large-scale standardized testing, individual administration as a means of assessing large numbers of examinees has virtually disappeared. Thus, since the 1920s the tests most commonly used in the educational arena have been designed for and administered to groups.

One exception to this, well known among language teachers, is the Language Proficiency Interview originally developed by the Foreign Service Institute of the U.S. Department of State (Clark, 1978; Frith, 1980). This test is individually administered and tailored to the examinee. Thus, if an examinee's initial responses to questions give evidence of high level proficiency in a second language, the examiner poses questions at a higher level of difficulty. In this manner each examinee receives a different test and one that is appropriate to his or her abilities.

Computerized adaptive testing (CAT) is analogous to the Language Proficiency Interview in that it tailors the test to the examinee. The technique involves using a computer to administer a test and to immediately analyze each response made by the examinee. The computer derives an early estimate of the examinee's ability level and then confirms or refines that estimate based on the examinee's response to subsequent items, which are selected according to the examinee's performance on previous items. This type of testing represents a major improvement over traditional group-administered tests. Traditional tests are of necessity constructed to be most precise for the student whose ability, or whose knowledge of the construct being measured, is about average. Thus, such tests contain a large number of items that are appropriate for the examinee with average or near-average performance, since these students are the most numerous given the typical normal distribution of scores. The farther a student is above or below average, the less appropriate is a group-administered test. On the other hand, given an adequate database a computerized adaptive testing program can produce tests that are equally appropriate to all examinees, regardless of their level of ability, achievement, or, in the case of second languages, proficiency.

Computerized adaptive testing involves more than just the use of a computer to deliver a test. A new approach to test and item analysis, generally referred to as either latent trait theory or item response theory, is involved also. Latent trait theory, as we shall call it initially for purposes of definition, assumes that there is a single underlying trait or ability that accounts for examinee performance on items. It provides a mathematical model that can be used to analyze an examinee's performance. The analysis focuses on the interaction between the ability level of the examinee and his or her performance on items with known parameters. These parameters always include an item's

difficulty and may include its discrimination power as well as the probability of guessing the correct answer. Although the mathematical foundation of latent trait theory was outlined in detail nearly two decades ago by Lord and Novick (1968), recent advances in technology are now making its practical application possible through computerized adaptive testing.

As indicated above, latent trait theory represents an innovative approach to test and item analysis. By comparison with traditional or classical test theory, it clearly seems to be superior. For example, under classical test theory, item statistics such as difficulty and discrimination power are heavily influenced by the sample of students that takes the test. A test given to a low-ability group will produce quite different item statistics than will the same test if given to a high-ability group. Thus, the item statistics depend on the group taking the test. On the other hand, with latent trait theory this is not the case. Under latent trait theory all test items are put on a common scale of difficulty. As a result, scores obtained on different tests can be placed on the same scale, even if the tests have no items in common and contain different numbers of items.

The discrimination index used in latent trait theory also seems superior. The discrimination power of an item analyzed under classical test theory, like the difficulty index, is also a function of the group taking the test. An item that is too easy or too difficult for the group taking the test may not show any useful discrimination power, yet the same item could be discriminating if presented to an appropriate group. The same can be said for the total test index known as reliability. Under latent trait theory the discrimination power of an item is calculated in terms of the ability level of the examinee. Thus, latent trait theory recognizes that the discrimination power of an item varies according to the ability level of the examinee.

Latent trait theory treats test reliability in an analogous manner. It permits one to determine the information function of a test for any given examinee. This information function is a measure of the test's precision of measurement for that examinee. Classical test theory, on the other hand, treats reliability as a constant.

This brief comparison of some of the features of both theories may provide the reader with a partial understanding of why language testing researchers across the world are becoming interested in latent trait theory. The articles appearing in Section One of this book should further illustrate the rationale for this trend and increase the reader's understanding of latent trait theory and computerized adaptive testing. In addition, it is hoped that these articles, as well as the articles in Section Two, will provide a fairly comprehensive overview of current activity involving the application of technology to the assessment of

language proficiency.

In the lead article in this collection, Peter Tung, using examples from his own test developed at the University of Hong Kong, illustrates how a computerized adaptive test is administered and scored. Tung then reviews major operational considerations in designing such a test. These considerations include a procedure for determining the difficulty of items to be presented before the computer arrives at an estimate of the examinee's ability level, the selection of subsequent items, the optimal length of the test, and the characteristics of the pool of items to be used by the computer. A third section of the article considers relevant test development matters, and an appendix defines some basic concepts and terminology used in response theory. While Tung provides a useful operational introduction to this new form of measurement, the subsequent articles in this section offer additional evidence of its potential.

Michael Canale discusses computerized adaptive testing of reading comprehension. After reviewing some basic literature in the field, he focuses attention on the need for research on the construct validity of reading comprehension tests and on the effects on the examinee's score of taking a test via computer.

Harold S. Madsen and Jerry W. Larson apply the Rasch model to the analysis of native language bias in three ESL tests. The Rasch model, the least sophisticated latent trait model, involves the analysis of a single parameter – item difficulty at different levels of ability. Their study demonstrates the application of this type of measurement to research on the role of examinee background variables in testing.

Grant Henning reports on the use of the Rasch model in the development of item banks for ESL placement tests using a database management program available for use with a wide variety of microcomputers. His technique can readily be applied to placement test programs at other institutions.

Marilyn M. Hicks uses item response theory to simulate a computerized adaptive placement test using TOEFL items. She establishes the validity of the test by correlating it with scores on the TOEFL.

John H.A.L. de Jong reports on the use of the Rasch model to develop group-administered tests of English, French, and German listening comprehension in the Netherlands. His program uses item pretest data to tailor tests to populations of examinees that differ in average ability level.

The articles in Section Two deal with other uses of technology in second language testing. Garry Molholt and Ari M. Pressler of the Speech Processing Laboratory of Rensselaer Polytechnic Institute describe their pioneering efforts to use a computer to rate the com-

prehensibility of second language learners' pronunciation of a portion of a standardized test of ESL. John L. D. Clark reports on his use of the language laboratory to test oral language proficiency in Chinese in a manner that accommodates standard rating scales. Winton H. Manning discusses a new type of cloze test and the innovative technology that he and his colleagues have developed to administer and score it. In the final article in this section, Joy Reid reports on initial studies of the use of microcomputer software developed at Bell Laboratories that she and others are using to teach and test ESL composition skills. Together, the four articles in Section Two demonstrate that computerized adaptive testing is only one of many possible applications of technology to language testing.

Although the articles in this book describe the activities of researchers, one can anticipate that advances made through these initial studies will impact soon on classroom language teachers. If technology can improve testing by making it more accurate or more efficient, this will make the teacher's job easier and curriculum will become more effective.

In pondering the not-too-distant future of second language testing, one can imagine that it will involve technology to a much greater degree than has been the case in the past. This technology could take many forms. It could include the use of complex new mathematical models of item analysis. Or, it may involve the use of highly sophisticated equipment that converts speech to electrical impulses and then measures those impulses against some standard. Other possibilities include an instantaneous quantitative and qualitative comparison of the second language learner's written language with that of educated natives. Surely, all these possibilities would be carried out with the aid of a computer. All these possibilities and others are exemplified in this volume. It is hoped that the reader will find it helpful and illustrative of things to come.

Charles W. Stansfield
August, 1985

References

Briere, E. J. & Hinofotis, F. B., (Eds.). (1979). *Concepts in language testing: some recent studies*. Washington, DC: Teachers of English to Speakers of Other Languages.

Clark, J. L. D. (Ed.). (1979). *Direct testing of speaking proficiency: theory and application*. Princeton, NJ: Educational Testing Service.

Frith, J. R. (Ed.). (1980). *Measuring spoken language proficiency*. Washington, DC: Georgetown University Press.

Jones, S., DesBrisay, M. & Paribakhp, T. (Eds.). (1985). *Proceedings of the fifth annual language testing colloquium*. Ottawa: Carleton University.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. New York: Addison Wesley.

Palmer, A. S., Groot, P. J. M. & Trosper, G. A. (Eds.). (1981). *The construct validation of tests of communicative competence*. Washington, DC: Teachers of English to Speakers of Other Languages.

SECTION ONE

New developments in measurement theory: computerized adaptive testing and the application of latent trait models to test and item analysis.

Computerized Adaptive Testing: Implications for Language Test Developers

Peter Tung
University of Hong Kong

Abstract

Computerized adaptive testing (CAT) is a new technology that promises to assess individual test taker's ability accurately and efficiently. This paper provides an introduction to CAT and the principles underlying its operation. Examples illustrating the procedure of this form of testing and the effect of varying various parameters associated with the test are given. The paper also considers the possibilities as well as limitations of CAT for language testing based on an examination of the requirements of CAT. It is argued that CAT is a viable technology for language testing, and that the use of CAT may prompt language test developers to formulate clearer descriptions of language proficiency constructs, resulting in a better understanding of the characteristics of test items.

Computerized Adaptive Testing: Implications for Language Test Developers

Accurate, efficient, individualized assessment of knowledge and skills utilizing high-speed electronic machines is on the horizon as a result of advances in mental measurement theories and the increasing use of computers in education. What does this development mean to the language teaching profession? This paper provides an introduction to a promising technology, computerized adaptive testing (CAT), and examines some of its implications for the evaluation of second language performance.

The principal feature of CAT is that test takers may receive different sets of test items tailored to the individual's ability, so that no one needs to put up with items that are too easy or too difficult. In general, by employing items that correspond to the ability of the test takers as nearly as possible, computerized adaptive tests are shorter than conventional fixed-length tests at the same level of measurement precision. Moreover, unlike conventional standardized tests, CAT programs based on Item Response Theory (IRT) and linked to an appropriate collection of test items can provide equally accurate ability estimates on a single scale for individuals representing a wide range of ability.

In the following section, an outline of the basic concepts of IRT employed in this paper is first given. Next, CAT is described with reference to an example illustrating how a test can be administered and scored while the test taker is responding to a series of test items at a computer terminal. Then, major considerations in the use of CAT systems are discussed: first, from the perspective of general test development, and second, with the requirements of language testing in mind. Several procedures for improving the efficiency and practicability of CAT will be illustrated.

The CAT program used in this study is based on a computer program for scoring computerized adaptive tests described in Bejar and Weiss (1979). It has been modified to run interactively on a VAX-11/750 computer and a SPERRY 1101/61 computer. A listing of the adapted source program can be obtained from this writer.

Some Basic IRT Concepts

Item Response Theory (IRT) covers a wide variety of models describing the probabilistic relationship between the response to a test item and the examinee's latent ability or the trait underlying the response. The simplest IRT models concern test items that measure a single trait and are scored as either correct or incorrect irrespective of the number of distractors presented to the test taker. The relationship between item response and examinee latent ability is then depicted by an Item Characteristic Curve (ICC), which has the general shape shown in Figure 1. The vertical axis in Figure 1 represents the probability of a correct response, which ranges in value from 0 to 1. The horizontal axis represents examinee ability, which usually ranges in value from -3.00 indicating low examinee ability to +3.00 indicating high examinee ability. It can be seen in Figure 1 that the probability of a correct response for a given item increases as the examinee ability improves.

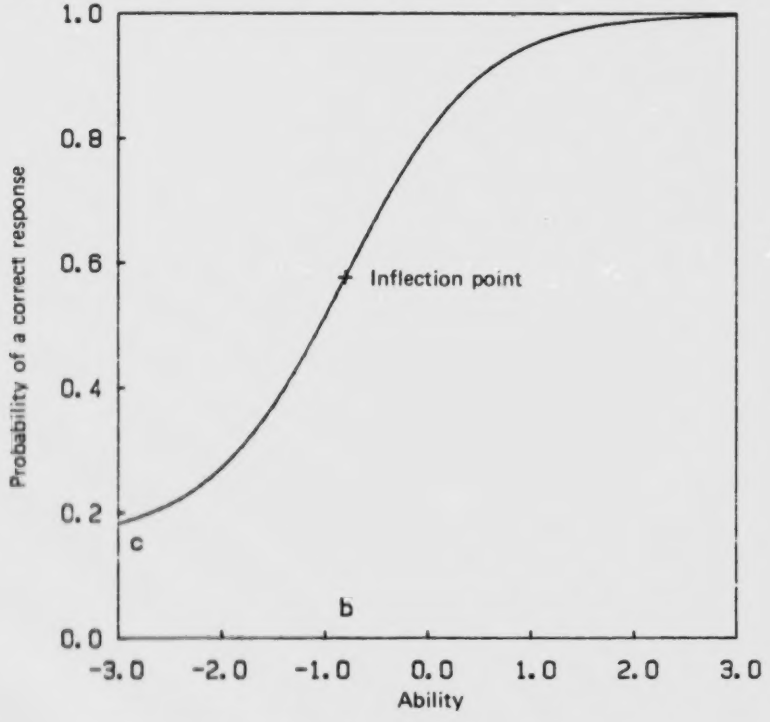


Figure 1. A typical ICC for a 3-parameter model of a dichotomously scored test item

An ICC may be described with three parameters. The "discrimination parameter" of the test item, customarily denoted by a, is proportional to the slope at the inflection point of the ICC (i.e., where the slope of the curve is maximum). That is, all other item parameters being equal, the steeper the slope of the ICC, the greater is the discriminating power of the test item in the region around the inflection point of the curve. In general, items with discrimination parameters of 1.0 or above are often considered satisfactory for the purpose of CAT. These items are relatively sensitive to small changes in examinee ability. Next, the "difficulty parameter" of a test item, denoted by b, is the value on the ability continuum corresponding to the inflection point of the ICC. A change in the difficulty parameter alone will result in a lateral shift of an ICC. Note that, in IRT, the test taker's ability and item difficulty are conveniently represented on the same scale. Lastly, the "guessing parameter", c, is the lower asymptotic value of the ICC. It gives the likelihood that an examinee with no knowledge of the subject matter will guess the correct answer of an item. Various procedures exist to estimate the values of the parameters for each item (see Ree, 1979, for a comparison of four computer programs for estimating item parameters). It is important to note that it is an empirical question whether an IRT model describes a test item adequately. Readers interested in a readable and comprehensive introduction to IRT may find Hulin, Drasgow, and Parsons (1983) helpful.

The Mechanics of CAT

As far as the test taker is concerned, a CAT session can be just a variation of the familiar paper-and-pencil test. The difference is that a video unit and a typewriter keyboard (or some more user-friendly devices) are used to display the questions and receive the answers. An interactive testing session may proceed as follows. The first message appearing on the screen, after the test supervisor has initiated the CAT program, may request the test taker to supply personal identification.

> Please enter your registration number below.
> Use the delete key to remove typing errors.
> When you have finished, push the return key.
> Type your registration number here (10 digits):

Provided that the test taker has entered a valid registration number, the first test item now appears on the monitor. As an illustration, a multiple-choice item is presented here.

Question 1:

Choose the best word or phrase to fill in the blank.

"George, you look tired. It's time you . . . a holiday."

(A) have
(B) had
(C) are having
(D) will have

Pleae enter your answer below.
Use the delete key to remove typing errors.
When you are sure of your answer, push the return key.
Type your answer here (one capital letter, A, B, C, or D):

More elaborate programs may incorporate such features as requests for verification of responses should unexpected answers be detected, and provide more detailed instructions if help is needed. If the test taker has followed the directions properly, subsequent test items will be presented until the CAT program arrives at a sufficiently accurate estimate of the individual's ability.

The last message in the testing session may then read:

You have just answered the last question.
Thank you for taking the test.
You will be informed of the results as soon as possible.

Unknown to the test-taker, however, the CAT program keeps track of the items attempted, estimates the individual's ability after each response, and selects the most appropriate questions to be administered. This series of functions can be explained with reference to Table 1, which comprises the beginning and final sections of the report of an imaginary examinee. Table 1 shows that the first item presented to the examinee was item No. 62 in the collection of items in an item pool. This item was selected in the absence of any prior information about the test taker. The number "1" after "Response = " indicates that the examinee answered the item correctly. An incorrect response would be designated by "0". The current program makes no distinction among the various incorrect responses.

Shown after the examinee response in Table 1 are three parameters characterizing the Item Characteristic Curve (ICC) of the test item. They are, in order, the discrimination parameter *a*, the item difficulty *b*, and the guessing parameter *c*. For the purpose of this example, the discrimination parameters of all the test items in the item pool were set equal to 1.50, indicating that the items could discriminate fairly well

Table 1
Portions of the CAT Report of an Examinee

Report of Examinee No. 1394882141

Item 1
No. 62 Response = 1 A = 1.50 B = 0.00 C = 0.25
Examinee has correct answers only
Ability estimate cannot be obtained with maximum likelihood procedure

Item 2
No. 82 Response = 1 A = 1.50 B = 1.00 C = 0.25
Examinee has correct answers only
Ability estimate cannot be obtained with maximum likelihood procedure

Item 3
No. 102 Response = 1 A = 1.50 B = 2.00 C = 0.25
Examinee has correct answers only
Ability estimate cannot be obtained with maximum likelihood procedure

Item 4
No. 122 Response = 0 A = 1.50 B = 3.00 C = 0.25
Ability estimate = 2.44 Standard error = 0.73
Best item next = 99 Item information = 4.58

Item 5
No. 99 Response = 1 A = 1.50 B = 1.90 C = 0.25
Ability estimate = 2.58 Standard error = 0.63
Best item next = 101 Item information = 4.60

Item 23
No. 106 Response = 0 A = 1.50 B = 2.20 C = 0.25
Ability estimate = 2.22 Standard error = 0.24
Best item next = 85 Item information = 3.38

Items answered = 23 Items not converged = 0

between the individuals who could answer them correctly and those who could not. Similarly, all the guessing parameters were arbitrarily set equal to 0.25, commonly taken to mean that there was one chance out of four for a test taker with absolutely no knowledge of the subject matter to guess the correct answer to an item. One practice in approximating the latter value is to construct items with four alternative answers, only one of which is correct.

In this idealized example, therefore, only the difficulty parameter varies in value from item to item. The item difficulty of the test items in the item pool ranges from -3.00 to +3.00 in steps of 0.10. It is assumed that two items were available at each difficulty level, making up a pool of 122 test items. To simplify our example further, let us assume that the imaginary examinee could answer correctly all items with item difficulty values below 2.05 and would answer incorrectly those items with difficulty values above 2.05. If items with a difficulty value of 2.05 had been available and administered, the examinee would have had a 50% chance of answering them correctly. As the test taker's ability and item difficulty are represented on the same scale in IRT, the last assumption means that the examinee had a "true" latent ability of about 2.32, after accounting for the guessing factor. (The value 2.32 is obtained using a standard equation describing the ICC of a test item. The mathematics need not concern us here.) The value 2.32 represents the ability that the CAT program tries to estimate using as few test items as possible. Table 1 shows that the first item presented to the examinee had a difficulty value of 0.0, so the examinee had little trouble coming up with the correct response. It may be noted that the simplifications above do not affect the generality of our description of CAT. The CAT program used here can readily handle test items with any valid parameters and the results in Table 1 would be the average outcome if the chance successes of the examinee had been simulated by a computer program.

The algorithm used in the CAT program to obtain an ability estimate from the examinee's pattern of correct and incorrect responses is known as the maximum likelihood estimation procedure, one of two frequently reported in the literature (see, for example, Weiss, 1982). This mathematical procedure, while relatively easy to implement, suffers from one serious drawback. Until the examinee's response pattern contains at least one right and one wrong answer, the procedure cannot produce an ability estimate. Consequently, in Table 1, as the examinee answered the first three items correctly (i.e., items No. 62, 82, and 102), no estimate of the examinee ability could be obtained, and this fact was reported accordingly. To make item administration possible for the first few items of the test, a mechanical item selection strategy was used. The mechanical rule was to increase the item difficulty in steps of 1.0 if the examinee answered the initial items correctly. Thus, the difficulty values of items 2, 3, and 4 were 1.0, 2.0 and 3.0 respectively. Similarly, if the examinee had answered the initial items incorrectly, the item difficulty values would have decreased in steps of 1.0. This mechanical procedure was used until the maximum likelihood procedure could take effect.

In Table 1, an estimate of the examinee's ability was obtained

mathematically after the fourth response. Associated with the ability estimate (2.44) is an indicator of its accuracy called the standard error of the ability estimate. In this case, the standard error of the ability estimate after four items is 0.73. An objective of adaptive testing is to reduce this standard error to a small value in the course of the test, thus ensuring the accuracy of the ability estimate. Following each ability estimate, the program selects from the remaining unused items in the item pool that item which, if answered next, would produce the highest amount of "information" about the test-taker calculated according to a standard formula. Table 1 shows that the best item to be administered after the fourth item was item No. 99, which produced an item information index equal to 4.58. The item was administered and scored, yielding a new ability estimate of 2.58. This estimate seems to be farther away from the true ability level of 2.32 assumed for our examinee than the estimate 2.44 obtained after the examinee's response to the fourth item. However, the standard error of estimate decreased to 0.63 from 0.73, indicating that a higher level of confidence can be placed on the new estimate. The cycle of choosing and displaying the next best item, and estimating the ability of the individual was then repeated.

For practical reasons, the length of the test must be limited. In this example, it was stipulated in the CAT program that the test-taker could not attempt more than thirty items in a single test. The test would also terminate automatically when the standard error of ability estimate decreased to a reasonably small value of 0.25. Table 1 indicates that after the 23rd item had been scored, the standard error attained a value below 0.25, and the test ended at that point. It is of interest to compare the final ability estimate produced by the CAT program with the postulated ability of the examinee. The final ability estimate obtained by the program was 2.22, which was well within one standard error of the true latent ability of 2.32. The latter value could only be obtained by the CAT program if an unlimited number of test items were available and administered.

General Considerations of CAT

The test developer making use of a CAT system should consider the following questions carefully. Which test item in an item pool is to be administered first by the CAT program? What are the rules governing the selection of subsequent test items for an individual? When does a CAT session terminate? What may be the number and the characteristics of items included in an item pool for CAT? While a subject specialist need not worry about fine technical details, a knowledge of the desirable features and the capabilities of CAT systems is essential in

order to determine the usefulness of various CAT programs and to exploit their numerous options.

The first consideration in the operation of a CAT system is the difficulty level at which a test-taker should begin a test. If no information about the test-taker is available, the best estimate of the individual's ability is at the mid-point of the ability scale, and an item with difficulty level 0.0 (or an item producing the maximum amount of information at the ability level 0.0) may be selected as the initial item. However, it is likely that some indication of an examinee's ability can be obtained before the test. Thus, if the group to which the individual belongs can be identified, or if the results of a related test are known, an item reflecting the mean ability of the group or the prediction of the previous test may be administered first.

Table 2
Selecting an Initial Test Item

Difficulty of first item[a]	Test length[b]	Final ability estimate[a]	Standard error[a]
0.0	23	2.22	0.24
1.0	22	2.22	0.24
2.0	21	2.22	0.24

Note. The test-taker's "true" ability is taken to be 2.32. The item pool characteristics are:
All the item discrimination parameters are set equal to 1.5. The item difficulty parameters vary between -3.0 and 3.0 in steps of 0.1, with two test items at each difficulty level (i.e., two items at -3.0, two items at -2.9, etc.), making an item pool of 122 items.
All the "guessing parameters" are set equal to 0.25.

[a]These values are measured on the same scale.

[b]This indicates the number of items in a test.

To illustrate that the efficiency of CAT can be improved if the ability of the examinee is taken into account in assigning the first test item, a series of computerized adaptive tests using identical item pools have been administered to the imaginary examinee, each test beginning with an item of different difficulty. Table 2 shows that the length of a computerized adaptive test increases as the difference between the difficulty level of the first item administered and the test taker's true

ability becomes greater, provided that the level of measurement accuracy remains constant. Computer simulations therefore indicate that when the difficulty level of the first item is chosen to be relatively close to the true ability of the individual, an accurate final estimate may be arrived at with fewer items, thus shortening the test. Hence the first desirable characteristic of a CAT system is that it should allow the test administrator to input relevant information about the test-taker, so that the test may be tailored to the individual right from the beginning (see also Weiss, 1982, p. 479; Green, Bock, Humphreys, Linn, & Reckase, 1982, p. 53).

Subsequently, mathematical methods are generally employed to select test items for presentation on the basis of some or all of the examinee's responses to the foregoing items. However, when examinee ability is not estimated mathematically, perhaps at the beginning of a test as in our example in the last section, mechanical item selection strategies may be used. The choice of a strategy or a combination of strategies should depend on a number of considerations. First, the strategies must not require enormous item pools to operate. This rules out mechanical strategies which neglect large sectors of an item pool during item selection and demand a fresh set of items for consideration every time an item is selected. Mathematical methods, on the other hand, make use of item pools much more efficiently and are compatible with small item pools. Second, the strategies must ensure that all the items within an item pool are optimally used. Even with mathematical methods, there is a danger that certain items with high discrimination indices may be selected for administration in nearly every test. If examinees can appear for examinations at different times, this constitutes a serious threat to test security. Various procedures have been proposed to ensure that in the long run, the selection of test items would not be heavily biased. Green et al. (1982, p. 52) have suggested that items might be chosen randomly out of the ten best possible items which could be administered at any instant. Third, the strategies must not take up so much computation time that item presentation is delayed, causing distraction and increasing test anxiety. Consequently, time-consuming procedures such as the maximum information item selection strategy described earlier may not be the most appropriate one for all applications, as in the case when microcomputers are used instead of large-scale computers. Moreover, it may be desirable to vary the strategies at different parts of a test depending on the nature of the skills assessed and the type of questions posed. (An example will be described in the next section.) Therefore, to meet the diverse needs of test development, a flexible CAT program ought to incorporate several item selection strategies and permit the test developer to choose among them according to the situation.

Another major consideration facing the user of a CAT system is the length of computerized adaptive tests. As computerized adaptive tests are tailored to the individuals, it is unnecessary for CAT programs to administer the same number of items to each test-taker as long as the final ability estimates are equally accurate. A survey of the results of various CAT programs reveals that when the standard error of the ability estimate is used as the criterion of measurement accuracy, a value of about 0.25 is a realistic goal for computerized adaptive tests to attain (Jensema, 1977; Lord, 1977; McBride, 1977; Samejima, 1977). However, there may be practical considerations which dictate that a computerized adaptive test should terminate before a desirable level of measurement accuracy is reached. For example, it may not be feasible for examinees to spend long hours in front of computer terminals, or to attempt more than a certain number of items in any one test. The quality and size of the item pool may also limit the number of suitable items for particular individuals. Such conditions for the early termination of computerized adaptive tests should be specified by the test developer using CAT programs, and if any of the conditions applies for an examinee, the condition as well as the standard error of ability estimate obtained should be reported together with the results.

The above manipulations of starting level, item selection strategies, and test length cannot compensate for the shortcomings of an item pool. At least four aspects must be considered in the construction of an item pool for CAT: dimensionality, the accuracy of the parameters characterizing the ICCs of the test items, the distribution of the parameters within the item pool, and the size of the item pool. CAT programs currently available assume that there is a single latent trait underlying the responses to all the items within an item pool. It is principally the item writer's responsibility to try to create test items measuring one theoretical construct, but because of the myriad ways in which test items may differ, no real item pools are likely to be strictly unidimensional. Various procedures exist to assess the dimensionality of an item pool (see, for example, Hulin, Drasgow, & Parsons, 1983, chap. 8), but as Lord (1980, p. 21) noted, there is no generally applicable statistical significance test to determine whether an item set is unidimensional. Consequently, Green et al. (1982, p. 26) suggested that more than one method might be employed to check the dimensionality of an item pool and if there is evidence for multidimensionality, subtests should be formed to assess each trait separately.

Before an item set can be utilized by a CAT program, the values of the parameters describing the Item Characteristic Curves of the items must be estimated. Both the item selection process and the estimation of a test-taker's ability depend greatly on the quality of the item

parameters. To ensure that the item parameter estimates approximate the unobserved true values of the parameters, it is necessary to collect the responses of a large number of examinees to each item. Current computer programs for estimating the item parameters require that each item be tried out with a minimum of 1,000 examinees. Furthermore, these examinees must represent all levels of test taker ability (Green et al., 1982, pp. 41-42). If the item pool contains too many items to be tried out on a single group of examinees at one time, different groups of individuals would be needed to obtain parameter estimates for all the items. A test developer would have to make sure that a valid design is followed to link the parameter estimates obtained from the different administrations (Green et al., 1982, pp. 47-48).

Table 3
Varying the Item Discrimination Parameter

Item discrimination parameter	Test length[a]	Test-taker's "true" ability[b]	Final ability estimate[b]	Standard error[b]
2.0	12	2.25	2.19	0.25
1.5	23	2.32	2.22	0.25
1.0	30	2.47	2.42	0.33

Note. The relevant item pool characteristics are:
The item difficulty parameters vary between -3.0 and 3.0 in steps of 0.1, with two test items at each difficulty level, making an item pool of 122 items.
The item "guessing parameters" are all set equal to 0.25.

[a]This indicates the number of items in a test.
[b]These values are measured on the same scale.

The next essential requirement of an item pool constructed for CAT is the distribution of the item parameters within the item pool. Repeatedly, studies have shown that in order to have reasonably short tests (less than 30 items), the values of the item discrimination parameters should exceed about 0.8 (Jensema, 1977; McBride, 1977; Samejima, 1977; Urry, 1977). Shorter tests will result if the values of the discrimination parameters are greater. Table 3 illustrates the effect of changing the average value of the discrimination parameters of the test items in an item pool: it would take many more test items to attain the same measurement accuracy if the average value of the discrimination parameters were lowered. To extrapolate the results a bit, it seems that the average value of the discrimination parameters needs to be around

1.3 in order for the desirable level of measurement accuracy (standard error ≤0.25) to be attained with less than thirty items in a test.

Although the effect of the guessing parameter is not as powerful, its value should be kept low. An item with a large guessing parameter will discriminate less well among examinees with extremely high (or extremely low) abilities. Urry (1977, p. 184) recommended that the values of the guessing parameters should be less than 0.3. There appears to be general agreement that the difficulty parameters of the test items should be evenly distributed throughout the ability range of potential test-takers, so that the maximum amount of information may be obtained about the test-takers by administering items of appropriate difficulty.

Table 4
Varying the Item Pool Size

Item pool size[a]	Final ability estimate[b]	Standard error[b]	Test length[a]
122	2.22	0.25	23
244	2.42	0.25	23

Note. The test-taker's "true" ability is taken to be 2.32. The item pool characteristics are: All the item discrimination parameters are set equal to 1.5. The item difficulty parameters vary in steps of 0.1, with the same number of test items at each difficulty level. All the "guessing parameters" are set equal to 0.25.

[a]Measured in number of test items.

[b]Measured on the same scale.

On the issue of the size of an item pool for CAT, recent studies seem to favor relatively small item pools, perhaps between one hundred and two hundred items (Jensema, 1977; Ree, 1981). Pioneering work on CAT by the staff of the Navy Personnel Research and Development Center has involved item pools ranging from 25 items to 225 items (Moreno, Wetzel, McBride & Weiss, 1984). Table 4 provides an illustration of the consequences of varying the size of item pools. It can be seen that doubling the item pool size has practically no effect on either the test length or the accuracy of the final ability estimates. However, the larger pool does allow the CAT program to select items yielding a higher information index after about 12 items have been

administered. Larger pools also tend to improve test security as the same items are less likely to be selected for different examinees. On balance, it seems that an item pool not much larger than 100 items of good quality is satisfactory for CAT.

Implications for Language Test Development

The implications of CAT for language test construction may not be difficult to see if we consider the process and requirements of this mode of testing. First of all, CAT assumes an accurately calibrated set of items that assess a single dimension of the examinee's ability. This presents the test developers with unique opportunities as well as problems. The requirement of unidimensionality means that language testers have to define their language proficiency constructs precisely and distinctly, and to ensure that the constructs are operationalized consistently across a wide ability range. But it also means that language testers will have the opportunity to verify their conception of the proficiency constructs which, for testing purposes, may be regarded as unidimensional. During the preparation of a set of test items for CAT, language testers can refine and develop the description of proficiency traits by observing the consequences on unidimensionality when certain items are included or excluded from the item pool. As the pretesting of the items involves a very large number of potential test-takers of a wide range of ability, the characteristics that make an item pool unidimensional may become clearer. Furthermore, language testers will learn in the same process what features contribute to the discrimination power, difficulty level, and susceptibility to guessing of items assessing a language trait. The insights derived from this exercise could supplement those obtained in conventional testing.

The drawback of the requirement of unidimensionality is obvious if everyday language behaviour cannot be adequately described on a single ability continuum. Frequently, a language event provides evidence for a number of components of a person's communicative competence such as grammar and sociolinguistic skills. This raises serious doubts whether complex language behaviour characteristic of real-life communication can be assessed with CAT.

However, it seems possible that CAT could be employed in diagnostic testing where language skills may be broken down into components such as knowledge of vocabulary items, knowledge of sentence structures, ability to perceive the appropriateness of utterances in particular contexts. It is more likely that such components will satisfy the demands of unidimensionality. This sort of work may also contribute to the development of sets of graded objectives for teaching and examining purposes.

Another implication of employing CAT in language testing is that the types of item formats that can be used may be limited. Until more advanced technology and computer software, such as an intelligent language parser, are available, item formats may be limited to multiple-choice items and items eliciting exact responses. Creative answers are not at present assessable by machines; nor are speaking skills likely to be judged successfully by computers in the near future.

It has been pointed out that the efficiency of a computerized adaptive test may be improved by matching as far as possible the test taker's ability with the difficulty level of the first item administered. In a language test, the effect of administering a first item that is too difficult may be much more serious than administering an item that is too simple for a given examinee. Whereas a test taker may treat an easy item lightly and make careless mistakes in answering it, a far too difficult test item might cause the test taker to become disheartened. This situation can be made worse if the test taker guesses the correct answer to the item and the CAT program proceeds to assign a more difficult item to the examinee. It may take the CAT program quite a few items to locate the approximate level of the test taker and to start assigning items of the right difficulty level. The situation may be better when a test taker chooses the incorrect answer of an easy item, but the examinee will still need to put up with answering some simple problems. It is essential therefore to adjust the difficulty level of the first item administered to the test taker based on prior information about the examinee. A simple routine can be easily built into the CAT program to allow the test administrator to adjust the difficult level of the first item.

For language testing, the usual mathematical item selection strategies in CAT programs may be too restrictive. It may be that several related questions need to be asked one after another to obtain a reasonable picture of an aspect of a test taker's competence. Consider the assessment of reading comprehension or some aspects of it which conform to the requirement of unidimensionality. It may be desirable to set more than one question on comprehension passages longer than a few sentences, either because it takes at least a couple of questions to determine the power of comprehension of the test taker with respect to a passage, or because it would be too tiresome for test takers to read more than a few passages in a test. In such cases, it is sensible to suspend normal item selection procedures and administer closely related items in sequence. Each reading passage in this example may be graded in difficulty based on the average value of the difficulty levels of the items associated with it. The mathematical item selection strategy may then be used to select passages rather than individual questions.

Similarly, the length of a language test need not be limited strictly to 30 items, even if the pre-set terminal level of measurement accuracy has not been reached after 30 items. Rather, it is an empirical question for the test developer to determine the maximum number of items of a particular type that may be administered to potential test takers without inducing fatigue or boredom in them.

Conclusion

CAT clearly has a number of important advantages over conventional paper-and-pencil, multiple-choice tests. First, it takes fewer test items for a computerized adaptive test to arrive at an accurate estimate of a test-taker's ability, even when the person's ability is near the extremes of the measurement range, that is, near -3.00 or +3.00. In contrast, conventional tests at best can match the accuracy of a computerized adaptive test for a very limited range of test-takers' ability as it is unfeasible for conventional tests to present test-takers with a large number of items assessing a wide range of ability. Second, the test items administered by a CAT program are determined on the basis of continual estimates of each examinee's ability. Items that are inappropriate for the examinees are not assigned. This may produce desirable affective effects on the examinees, who may discover that the test items are always challenging but never much beyond their capability. Third, as each test taker essentially takes a different test, test security is enhanced. That the number of items in an item pool need not be very large is an added bonus to the test developer. A case may therefore be put forward that CAT is a viable technology for language testing, provided that its limitations are recognized and steps taken to adapt it to meet the special needs of language test development.

References

Bejar, I. I., & Weiss, D. (1979). Computer programs for scoring test data with item characteristic curve models. *JSAS Catalog of Selected Documents in Psychology, 9*(3), 49.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1982). *Evaluation plan for the computerized adaptive vocational aptitude battery* (Research Report 82-1). Baltimore, MD: The Johns Hopkins University, Department of Psychology.

Hulin, C. L., Drasgow, F., & Parsons, C. K., (1983). *Item response theory: application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Jensema, C. J. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement, 1*(1), 111-120.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*(1), 95-100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

McBride, J. R. (1977). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1*(1), 121-140.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8*(2), 155-163.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3*(3), 371-385.

Ree, M. J. (1981). The effects of item calibration sample size and item pool size on adaptive testing. *Applied Psychological Measurement, 5*(1), 11-19.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*(2), 233-247.

Urry, V. W. (1977). Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement, 14*(2), 181-196.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492.

The Promise and Threat of Computerized Adaptive Assessment of Reading Comprehension

Michael Canale
The Ontario Institute for Studies in Education, Toronto

Abstract

This paper examines the potential of computerized adaptive testing (CAT) with respect to two fundamental issues in language testing research: construct validity and test method effects.

With respect to the issue of the construct validity of assessment of reading comprehension, the most important potential strength and weakness of CAT involve its heavy dependence on the model of measurement theory often referred to as item response theory (or more generally as latent structure analysis). In particular, the main strength of such a version of CAT is that it must be based upon a sound and explicit analysis of different ability levels in reading comprehension – i.e., upon a sound construct or theory of reading comprehension. Perhaps the main weakness of this version of CAT is that the construct to be measured must, according to item response theory, be unidimensional – i.e., largely involve only one factor. Not only is it difficult to maintain that reading comprehension is a unidimensional construct (for example, to ignore the influence of world knowledge), but it is also difficult to understand how CAT could serve useful diagnostic and achievement purposes if reading comprehension is assumed to be unidimensional and, hence, neither decomposable into meaningful subparts nor influenced by instruction.

With respect to the issue of test method effects in assessment of reading comprehension, CAT offers a unique strength and weakness. The main strength may be the use of assessment techniques that are not easily administered in a non-CAT format. The main weakness may be the tendency among both large-scale and small-scale test developers simply to mechanize exiting product-oriented reading comprehension item types in an effort to enhance the efficiency of test administration and assure conformity to crucial assumptions of item response theory.

The Promise and Threat of Computerized Adaptive Assessment of Reading Comprehension

In the conclusion to his lead article in the special issue of *Scientific American* (September, 1984) devoted to computer software, Alan Kay, one of the most insightful pioneers in computer science, focuses on the nature of computer literacy by asking the following: Is the computer simply a car to be driven or rather is it an essay to be written? That is, should we view the computer as just another tool or appliance that is simply to be mastered and put to use, or should we view it as a language or medium that is capable of generating unanticipated goals and images that may permit us "to encounter the very structure and style of thought and imagination" (p. 59)?

Some of the promises and threats implicit in Kay's question on computer literacy are more explicitly stated in a recent article on computers and language learning written by Frank Smith, one of the most insightful pioneers in research on reading and literacy. Smith (1983) registers concern that language educators will simply use the microcomputer to mechanize and control current classroom practice rather than to enhance and explore learners' encounter – and inherent fascination – with language, literacy and learning.

The tone of the present paper reflects that of Kay's article: the theme of this volume – technology and language testing – is still very much an essay to be written. The title of my paper reflects that of Smith's article, both to signal that my views about the use of technology in language testing are similar to his views about the use of technology in language learning and to signal that my goal here is not to examine hard data nor to draw empirically based conclusions. The general goal of this paper is rather to examine the potential of computerized adaptive testing (CAT), i.e. the use of the computer to tailor a test to a test-taker's ability as the test is actually administered. In particular, CAT is examined with respect to two fundamental and related issues in language testing research: construct validity and test method effects. To this end my focus is on certain major strengths and weaknesses of computerized adaptive techniques for assessing one skill area of language proficiency, reading comprehension skills. The choice of reading comprehension is appropriate given the importance and complexity of this skill area as well as the fact that it is the most frequently addressed component of language proficiency in research on CAT.

Construct Validity, Reading Comprehension and CAT

As understood here, the general notion of construct validity in testing refers roughly to how well we can interpret the data we create with measurement instruments in terms of some theory – or construct – of a given domain or, put another way, how well we understand what we are and should be measuring. Educational measurement experts such as Messick (1981) have argued that the two most important considerations in psychological and educational testing are construct validity and the potential social consequences of test data. Without downplaying in any way the importance of the ethical use of tests, our focus on construct validity is certainly warranted in the area of reading comprehension tests. This is because our constructs of reading comprehension are still viewed by most experts in reading theory and reading assessment as tentative and limited at best (see, for example, Curtis & Glaser, 1983).

With respect to this major issue of the construct validity of current tests of reading comprehension, perhaps the most important potential strength and weakness of CAT involve its currently heavy dependence on the model of measurement theory often referred to as item response theory (or more generally as latent structure analysis) elaborated by Lord (1980) and others. A few of the essential aspects of item response theory will be clarified in the discussion of promises and threats below.

The Promise.

The main promise of CAT for addressing the construct validity issue may appear ironic: CAT will require us to develop sound and explicit theories for the domains to be measured. Thus, for example, adequate constructs of reading comprehension are a prerequisite for any CAT that is worthwhile in precision and efficiency of measurement for reading comprehension skills. This is so because item response theory, upon which current CAT is largely based, is a mathematical model that directly relates the level of difficulty of a given task (or item) to a given ability level within a domain. The sole basis of this relationship – and hence the crucial element in both item/test design and data interpretation in CAT – is our theory of the domain itself. Without such a theory, it is difficult to argue that the item banks required for CAT provide complete and consistent coverage of a domain (as they must under item response theory) or that we have rejected and accepted items in the bank for the right reasons (see McDonald, 1981, on this last point).

It is important to emphasize that this potential promise of CAT is also possible, but not necessarily guaranteed, in classical test theory. This is because classical test theory – the measurement theory underlying most large-scale, norm-referenced educational testing – does not re-

quire explicit and absolute ability scale descriptions for the domain in question. Notions such as the level of difficulty of a task are defined only relative to such specifics of the testing occasion as the other items on the test and the other individuals taking the test (see Kreitzberg, Stocking, & Swanson, 1978; Wainer, 1983; Weiss, 1983). It is perhaps partially for this reason that educational measurement research during the past fifty years has not, according to many of the foremost educational measurement experts, led to development of adequate descriptive frameworks for educational domains such as reading comprehension (see especially Shoemaker, 1980, for discussion).

Some educational measurement research has, nonetheless, contributed directly to development of such theoretical frameworks. A great deal of this research has been undertaken within the recent context of criterion-referenced testing rather than within the more established context of norm-referenced testing. Whereas a norm-referenced test consists of tasks designed to maximize differences in performance among test takers, a criterion-referenced test consists of whatever tasks (or sample of tasks) test takers must actually be able to perform for a given purpose and at a given ability level (see Popham, 1978). In the context of the present paper, perhaps the most relevant and instructive work in criterion-referenced testing is that represented in the collaborative effort of the Interagency Language Roundtable of the U.S. Government, the American Council on the Teaching of Foreign Languages, and Educational Testing Service. Table 1 below broadly outlines a part of the ability level descriptions for reading and listening comprehension that these groups have developed – in a preliminary form – based on their joint experience over many years and in many different language contexts (see Lowe, 1985). While these descriptions have still to be empirically validated and in no way constitute a theory, they nonetheless represent the type of absolute ability scale that is required to support CAT of reading comprehension. In sum, such ability level descriptions provide important building blocks for construction of the more rigorous, process-oriented and complete theories of reading comprehension that CAT will require us to develop.

The Threat.

For construct validity purposes, perhaps the main threat of current versions of CAT based on item response theory is that CAT may – in direct conflict with the promise above – require us to trivialize our theories of reading comprehension and to compromise their relevance to educational settings. This is because current computer implementations of item response theory require that the construct to be measured be unidimensional, i.e. be assumed largely to involve only one major factor or underlying trait. Thus, for example, to implement

Table 1
Text and Task Specifications for Reading (R) and Listening (L)

Level	Text type	R/L function	R/L strategy
0/0+	Enumerative (numbers, names, street signs, isolated words/phrases)	Discern discrete elements	Recognize memorized elements
1	Orientational (simplest connected text, such as concrete descriptions)	Identify main idea	Skim, scan
2	Instructive (simple authentic text in familiar contexts and in predictable sequence)	Understand facts, literal information	Decode, classify
3	Evaluative (authentic text on unfamiliar topics)	Grasp ideas and implications	Infer, guess, hypothesize, interpret
4	Projective (all styles and forms of text for professional needs or for general public)	Deal with unpredictability and cultural references	Analyze, verify, extend hypotheses
5	Special purposes (extremely difficult and abstract text)	Equivalent to educated native reader/listener	All those used by educated native reader/listener

current versions of CAT of reading comprehension one must assume that reading comprehension is largely a unidimensional construct.

Such an assumption threatens to be trivializing and compromising in the following senses. First, it is overly reductionist and misleading to maintain that reading comprehension comprises only one major dimension, whatever that dimension may be. Such an assumption would require us to ignore the already well documented influence of such variables as world knowledge (e.g. Cummins, 1981), cultural background (e.g. Covey, 1983; Irvine & Berry, 1983), language background (e.g. Hoover & Politzer, 1985), sex differences (e.g. Donlan, Ekstrom, Lockheed & Harris, 1977), literacy in the home (e.g. Wells, 1981), general reasoning skills (e.g. Singer, 1977), the type of text (e.g. Child,

1985; Moirand, 1979), and reading styles (e.g. Bussis, Chittenden, Amarel & Klausner, 1984). While one could no doubt empirically derive unidimensional theories of reading comprehension (of which there would be many) and use such theories to support CAT, it seems perverse to do so – almost like trying to make the foot fit the shoe. In the end, such an endeavor is more likely to hinder than to further our understanding of the construct of reading comprehension.

Second, and more generally, the assumption of unidimensionality threatens to compromise the value of CAT for educational achievement and diagnostic purposes. It is difficult to understand how CAT could serve useful achievement and diagnostic purposes if reading comprehension, for example, is assumed to be unidimensional and consequently neither influenced by instruction nor decomposable into meaningful subparts. This general threat is stated well by Traub and Wolfe (1981).

> Many educators, in their eagerness to apply latent trait theories, have neglected to question their reasons for doing so. They have neglected to ask whether the assumptions underlying these theories, assumptions implying that examinee responses to test items are free from the influence of instruction and all other contextual factors, are realistic . . . Is it sensible to talk about "latent traits" of educational achievement? Is it likely that all aspects of achievement in a subject can be captured in a single latent trait? (pp. 377-379)

Traub and Wolfe shift their emphasis to content validity – here the relevance of a test for a given educational context – as follows.

> Suffice it to say here, there exists no prescription for achievement test development with which we are familiar that recommends the use of statistical information (e.g. fit of items to a latent trait model) as a means of assessing the content validity of a collection of items . . . If the tester's purpose is to assess student achievement of realistic domains of content, the kind that students and teachers deal with on a daily basis, then fit to a unidimensional latent trait model seems a singularly inappropriate criterion to use in judging content validity: this criterion will almost certainly restrict the range of content that is tested. (pp. 382-383)

A similar argument is advanced by J. B. Carroll (1983) in his critique of unitary (or unidimensional) models of language proficiency for instructional purposes.

The specific and general threats posed by the unidimensionality requirement have not, of course, gone unnoticed in the literature on applications of CAT – unidimensionality is one of the most debated and actively researched issues. Several suggestions have been proposed for addressing this problem. One is to continue to explore multidimensional versions of item response theory which, according to Traub and Lam (1985), are possible in principle but are extremely difficult to implement in practice without severely limiting CAT. Another is to explore the use of item "clusters" or CAT subtests within larger tests, where each subtest would be assumed to measure a unidimensional trait (e.g. reading comprehension on a particular topic). This suggestion is discussed by Green, Bock, Humphreys, Linn and Reckase (1984). A point not clear in this suggestion is whether and how the larger test could be administered in a CAT format unless a model analogous to a multidimensional version of item response theory were used. One final suggestion emerges from recent work on proficiency-oriented language testing in educational contexts, such as the work represented in Table 1 above (see Lowe, 1985). Such proficiency-oriented testing focuses on criterion tasks that characterize actual language performance at different ability – or proficiency – levels. It has been suggested that such proficiency-oriented testing may be more appropriate than traditional achievement and diagnostic testing for those language classrooms in which ability to use language, rather than simply knowledge about language, is the primary goal (e.g. Clark, 1983; Omaggio, 1983). It is thus possible that proficiency-oriented testing of reading comprehension may provide a more suitable and relevant application of CAT to educational assessment than the more traditional achievement and diagnostic orientations that Traub and Wolfe (1981) have questioned.

Test Method Effects, Reading Comprehension and CAT

While not generally acknowledged to be as fundamental an issue in testing research as is construct validity, the effect of test methods on test scores is considered as a fundamental issue in this paper for two reasons. First, the analysis of test method effects is closely related to that of construct validity. Just as an inadequate understanding of what we are and should be measuring limits the construct (and content) validity of any test, so does our inadequate understanding of how different test methods (e.g. multiple choice, long constructed response) influence performance also limit test validity. Elaborate experimental designs have been proposed to address such method effects: for example, multitrait and multimethod validation techniques such as those employed by Bachman and Palmer (1981, in press) for

analyzing various tests of reading and speaking. As well, careful studies have been undertaken to investigate the varying influence on test scores of a test taker's attitudes, understanding and training with respect to formal testing and particular test methods (e.g. Cohen, 1984; Deyhle, 1983). The second reason for examining test method effects in this paper is that CAT introduces a largely new method and medium into the assessment of language skills. It would be naive to assume that the combined novelty of adaptive testing and the computer does not call for careful consideration of possible sources of method effects in the assessment of reading comprehension. It is argued below that CAT does in fact offer at least one unique promise and threat.

The Promise.

The most unique promise of CAT for reading comprehension testing may be the potential for use of highly desirable assessment techniques that are not easily administered in a non-CAT format. Such techniques could permit more process oriented, continuous, interactive and unintrusive language tasks that can be naturally integrated into purposeful and thematically coherent language use and language learning activities in a classroom.

Many suggestions for such techniques have surfaced in work on CAT. These suggestions range from constructed responses, scaled or probabilistic responses, simulations, interactive video, group collaboration via networking, to optical scanners and voice input devices (see, for example, Baker, 1984; Kreitzberg, Stocking & Swanson, 1978). Perhaps the most exciting prospects come from outside the field of educational measurement: those from the field of artificial intelligence. Especially promising are the techniques currently under development and study in what are sometimes called "intelligent tutoring systems." Venezky (1983) describes such computer-assisted activities as follows:

> Programs such as [these] have in common the capability to build a model of the students' abilities, a representation of expert behavior in relation to the skills to be taught, and a representation of the subject matter itself. Through diagnosis of the students' strategies and their relationship to expert strategies, instruction is generated . . . . There are no predetermined, frame-by-frame sequences, nor are there fixed responses to be given to particular student answers. (p.41)

One of the best examples of such a system is a math game (for the PLATO computer system) entitled "How the west was won" (Burton &

Brown, 1982). This system is of interest for several reasons: it is process-oriented rather than product-oriented; it is intended to diagnose and construct a model of the learner's strengths and weaknesses in an unintrusive manner; it has a game format; and it adapts the game to the learner's level of performance while still seeking to "coach" or offer the learner assistance to find out both how much his or her abilities and performance can be improved as well as what types of assistance are effective. Learners are relatively free to use or ignore the assistance offered.

The educational philosophy behind such adaptive coaching environments has been articulated and explored by Papert (1980) and Burton, Brown and Fischer (1984), among others. As models for CAT, such environments are of interest because they provide images of educational assessment that are not often explored for practical, and no doubt other reasons. At the same time, such models are problematic for CAT in that the adaptation of tasks is not based on any measurement theory as carefully articulated and researched as is item response theory.

Yet work on such "intelligent tutoring systems" is promising for CAT of reading comprehension. Computer-simulation studies of reading processes have become increasingly informative and interesting (e.g. Shank, 1984), and unintrusive, interactive techniques such as oral protocol analyses have been profitably used both on and off computer to study expert and nonexpert reading strategies (e.g. Caes & Bereiter, 1984; Fillmore, 1981). Such research and images provide promising stepping stones if we are interested in moving toward more learner controlled, process-oriented and unintrusive assessment events in the language classroom (see Canale, 1985, for discussion).

The Threat.

The unique threat of CAT as concerns method effects is that it facilitates the mechanization of existing product-oriented tests of reading comprehension – the mechanization of both what existing tests include and what they represent. For example, banks of multiple-choice items already assembled, or easily assembled within existing test development plans, can no doubt be automated via CAT to deliver an efficient, precise, product-oriented and thematically incoherent mix of short reading passages each followed by a question or two.

Such automation is perhaps a useful first step in our study of the extent to which CAT may introduce method effects different from those in our traditional tests of reading comprehension (see Green et al., 1984, for discussion). But the threat is partially that CAT in reading comprehension may not move beyond this first step; both large-scale and small-scale test developers may expend most of their energy and limited resources on ways to enhance the efficiency of test administra-

tion and to ensure conformity to crucial assumptions of current item response theory. Even without the added complications of CAT it is difficult to effect change in reading comprehension testing. Chittenden (1983) concludes his study of the effects of traditional reading test items on certain test takers as follows:

> In general, the gap between our research knowledge of children's learning and our tests for assessing that learning continues to widen. The proceedings of the National Conference on Research on Testing (. . . 1979) point up this discrepancy. While the cognitive sciences have added substantially to our understanding of reading, of mathematics learning, etc., the paradigm for achievement testing, constructed in the 1920s, has remained essentially unaltered (Buros 1977). The paradigm is well suited to sequential models of learning, but poorly responsive to multifaceted accounts and to the fact that individual children may follow different routes to comparable levels of attainment. For both theoretical and practical reasons, the field of reading assessment appears especially overdue for development of new measures. The present analysis of implications of stylistic differences in children's learning draws attention to a dimension that is neglected in much of the current reappraisal of testing. (p. 42)

It is this threat that points so clearly to the general threat that both Allen Kay (1984) and Frank Smith (1983) identify: the threat of social and educational control that the computer – and certainly CAT – can serve only too well. No one has written more persuasively on this threat than Marianne Amarel (1983) of Educational Testing Service. She writes:

> The design of curricula, the setting of objectives, and the evaluation of achievement are taking place at some distance from the classroom, accomplished by subject-matter specialists, publishers, and district or state administrators. Necessarily, and perhaps intentionally, these trends curtail the exercise of teachers' judgments and diminish their decision-making responsibilities. Planning, designing, and evaluating educative experiences, activities considered integral to the teaching role, are passing from the teachers' domain to that of administrators who formulate these decisions in positions well removed from the place where teaching and learning occur. Some view these changes as an

> attempt to undermine the already fragile professional status of the teacher, to be designed to "de-skill" the classroom teacher in order to mechanize and thereby gain firmer control over the conduct of education (Apple 1982). The interests and ideologies that fuel these trends may be disputed, but their direction is generally acknowledged. (p. 25)

Concluding Comments

Promises and threats do not provide the data needed to draw serious conclusions about a complex and consequential theme such as technology in language testing or the topic of computerized adaptive assessment of reading comprehension. Moreover, the promises and threats identified in this paper are only a few drawn from a larger and perhaps equally important set. In spite of these limitations, three suggestions are nonetheless offered below.

First and foremost, we would do well to reconsider our current image of reading comprehension testing – with or without CAT – in light of the "intelligent tutoring systems" and simulation environments being explored in research in artificial intelligence. More learner controlled, process-oriented and unintrusive adaptive assessment techniques are already in experimental use in such research. It would be interesting to study just how much precision and efficiency of measurement is actually lost in these adaptive environments where item response theory is not used as a basis for design and delivery of tasks. It would also be interesting to study users' attitude toward CAT driven by item response theory, on the one hand, and "intelligent" tutoring environments on the other.

Second, two types of content validity studies would be useful. One would address current ability scales for reading comprehension, such as those sketched in Table 1. The main purpose of such a study would be to flesh out the range of tasks at each level (horizontally) in order to construct a more fine-grained continuous interval scale of reading comprehension abilities, both within levels and from level to level (vertically). The current reading proficiency scale reflected in Table 1 is neither an equal interval scale nor a complete ability scale (see Lowe, 1985). As such, it is suggestive but inadequate as a basis for CAT of reading comprehension. The other type of content validity study would examine the extent to which, for the same testing purpose and context, the content of proficiency-oriented reading comprehension tests constructed without reference to item response theory is similar to the content of reading comprehension tests constructed on the basis of item response theory. The goal of such studies would be to identify just how much item response theory does in fact limit or exclude

otherwise desirable content (such as tasks, topics, types and lengths of text). Such studies could also serve as a basis for assembling item clusters or subtests that could be administered via CAT.

Finally, of course, there is need for studies on alternative and refinements to the measurement models on which CAT is currently based. An excellent overview of current studies and future directions is provided by Traub and Lam (1985). Particularly interesting is their analysis of item sampling models, which serve to estimate the proportion of tasks in a domain that an individual can perform successfully based on his or her performance on a random sample of items. Such item sampling models are not adaptive and do not generate data on the value of different tasks for correctly estimating an individual's precise ability level. However, for our purposes such models are worth considering since they do not require the assumption of unidimensionality, are currently implementable on computer, and could be useful for testing within ability levels (horizontally) to address the question of how many of the various multidimensional tasks at a given level someone can actually perform.

In short, there are still many, many essays to be written on reading comprehension, language testing, and the role of the computer in education. As we try to write them, we may be reminded of the words of W. S. Merwin at the close of his poem "An encampment at morning":

> like the shapes of snowflakes
> we are words on a journey
> not the inscriptions of a settled people

References

Amarel, M. (1983). The classroom: an instructional setting for teachers, students and the computer. In A. C. Wilkinson (Ed.), *Classroom computers and cognitive science*. New York: Academic Press.

Apple, M. W. (1982). Reproduction and contradiction in education. In M. W. Apple (Ed.), *Cultural and economic reproduction in education: essays on class*. Boston: Rutledge & Kegan Paul.

Bachman, L., & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. Palmer, P. Groot & G. Trosper (Eds.), *The construct validation of tests of communicative competence*. Washington, DC: Teachers of English to Speakers of Other Languages.

Bachman, L., & Palmer, A. S. (in press). *Basic concerns in language test validation*. Reading, MA: Addison-Wesley.

Baker, F. B. (1984). Technology and testing: state of the art and trends for the future [Special Issue]. *Journal of Educational Measurement, 21*(4), 399-406.

Buros, O. K. (1977). Fifty years in testing: some reminiscences, criticisms and suggestions. *Educational Researcher, 6*, 9-15.

Burton, R. R., & Brown, J. S. (1982). An investigation of computer coaching for informal learning activities. In D. H. Sleeman & J. S. Brown (Eds.), 1982, *Intelligent tutoring systems*. London: Academic Press.

Burton, R. R., Brown, J. S., & Fischer, G. (1984). Skiing as a model of instruction. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: its development in social context*. Cambridge, MA: Harvard University Press.

Bussis, A. B., Chittenden, E. A., Amarel, M., & Klausner, E. (1984). *Inquiry into meaning: an investigation of learning to read*. Hillsdale, NJ: Erlbaum.

Canale, M. (1985). *Proficiency-oriented achievement testing*. Paper presented at the 1985 ACTFL Master Lecture Series. Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research*. Englewood Cliffs, NJ: Newbury House.

Case, R., & Bereiter, C. (1984). *From behaviourism to cognitive behaviourism to cognitive development: steps in the evolution of instructional design*. Toronto, Ontario Institute for Studies in Education, Centre for Applied Cognitive Science.

Child, J. R. (1985). Language proficiency levels and the typology of texts. In J. R. Frith & J. E. Liskin-Gasparro (Eds.), *Measuring spoken language proficiency* (Vol. 2). Washington, DC: Georgetown University Press.

Chittenden, E. A. (1983). *Styles, reading strategies and test performance: a follow-up study of beginning readers* (ETS Research Report 83-81). Princeton, NJ: Educational Testing Service.

Clark, J. L. D. (1983). Language testing: past and current status, directions for the future. *Modern Language Journal, 67*(4), 431-43.

Cohen, A. D. (1984). On taking language tests: what the students report. *Language Testing, 1*(1), 70-81.

Covey, M. (1983). *Cultural conventions and expectations of two cultural groups in reading and writing*. Unpublished master's thesis, University of Toronto, Ontario.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: a reassessment. *Applied Linguistics, 2*, 132-149.

Curtis, M., & Glaser, R. (1983). Reading theory and the assessment of reading achievement. *Journal of Educational Measurement, 20*, 133-148.

Deyhle, D. (1983). Between games and failure: a micro-ethnographic study of Navajo students and testing. *Curriculum Inquiry, 13*(3), 347-376.

Donlan, T. F., Ekstrom, R. B., Lockheed, M. E., & Harris, A. (1977). *Performance consequences of sex bias in the content of major achievement test batteries* (ETS final project report). Princeton, NJ: Educational Testing Service.

Fillmore, C. J. (1981). Ideal readers and real readers. In D. Tannen (Ed.), *Georgetown University Roundtable on Language and Linguistics 1981*. Washington, DC: Georgetown University Press.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

Hoover, M. R., & Politzer, R. (1985). Bias in reading tests: a linguistically oriented view. Manuscript submitted for publication.

Irvine, S. H., & Berry, J. W. (Eds.). (1983). *Human assessment and cultural factors*. New York: Plenum Press.

Kay, A. (1984, September). Computer software. *Scientific American*, pp. 52-59.

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. (1978). Computerized adaptive testing: principles and directions. *Computers & Education, 2*, 319-329.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lowe, P. (1985). The ILR Proficiency Scale as a synthesizing research principle: the view from the mountain. In C. J. James (Ed.), *Foreign language proficiency in the classroom and beyond*. Lincolnwood, IL: National Textbook.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

Mesick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20.

Moirand, S. (1979). *Situations d'écrit (compréhension, production en langue étrangère)*. Paris: Community Languages and Education International.

Omaggio, A. C. (1983). *Proficiency-oriented classroom testing*. Washington, DC: Center for Applied Linguistics.

Papert, S. (1980). *Mindstorms: children, computers and powerful ideas*. New York: Basic Books.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Shank, R. W. (1983). *The cognitive computer*. Reading, MA: Addison-Wesley.

Shoemaker, D. M. (1980). Improving achievement testing. *Educational Evaluation and Policy Analysis, 2*, 37-49.

Singer, H. (1977). IQ is and is not related to reading. In S. M. Wanat (Ed.), *Issues in evaluating reading*. Washington, DC: Center for Applied Linguistics.

Smith, F. (1983). The promise and threat of microcomputers in language learning. In J. Handscombe, R. A. Orem & B. P. Taylor (Eds.), *On TESOL '83: the question of control*. Washington, DC: Teachers of English to Speakers of Other Languages.

Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36*, 19-48.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), *Review of Research in Education, 9*, 377-435.

Venezky, R. (1983). Evaluating computer-assisted instruction on its own terms. In A. C. Wilkinson (Ed.), *Classroom computers and cognitive science*. New York: Academic Press.

Wainer, H. (1983). On item response theory and computerized adaptive tests. *Journal of College Admissions, 28*, (4), 9-16.

Weiss, D. J. (1983). Introduction. In D. J. Weiss (Ed.), *New horizons in testing: latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Wells, G. (1981). *Learning through interaction: the development of language*. Cambridge: Cambridge University Press.

Author's Notes

This paper is dedicated to Marilou Covey, who was killed in the Korean Airline Flight 007 massacre in 1983. Invaluable assistance in the preparation of this paper was provided by my colleagues Graham Barker and Ross Traub as well as Charles Stansfield of ETS. I am grateful to them.

Computerized Rasch Analysis of Item Bias in ESL Tests

Harold S. Madsen and Jerry W. Larson
Brigham Young University

Abstract

Test bias is a recurring concern for second language testers. It appears in many forms, from cultural and socio-economic bias to bias that results from anxiety about having to take certain types of tests. This study uses Rasch-generated test-item analyses to assess bias in three multiple-choice ESL subtests – grammar, listening comprehension and reading comprehension. Two hypotheses are tested: (1) There are identifiable patterns of items that generate a poor fit statistic, which can be accounted for by student language background. (2) Similarly, there are identifiable patterns of items that produce a poor fit statistic, regardless of language background. The data collected in the study are Rasch-analyzed using a software package written for micro-computers. A brief description of this program, its functions and data printouts is given. Results of the analysis do not reveal apparent item bias related to language background, but the subtests are apparently biased in favor of intermediate ability students.

Computerized Rasch Analysis of Item Bias in ESL Tests

The study reported in this paper applies a computer software application of latent trait measurement to identify item bias in ESL subtests.

Contemporary item response theory manifests itself in various statistical models under the rubric of Latent Trait Analysis. The model that seems most appropriate for relatively small-scale language test analysis (utilizing an N of only 100, and sometimes even less) is the one-parameter Rasch model, which focuses on item difficulty and student ability. The Rasch procedure has been successfully utilized for such varied purposes as precise test-item calibration (Englehard, 1980), learner-specific item analysis (Rasch, 1966, pp. 49-57), item banking (Rubin & Mott, 1983), test linking, and computerized testing (Henning, 1984, pp. 123-133). The Rasch model can also assess the precision with which a test measures individuals at a variety of ability levels, through the calculation of an "information curve" (Hambleton & Cook 1977, pp. 75-96).

Recent software developed expressly to perform the Rasch analysis on a microcomputer makes it feasible to conduct analyses of item *bias* almost as easily as preparing the traditional test-item analysis, though it is limited by some of the same constraints of the item analysis. (For example, poor discrimination or weak distractors can be identified through routine, mechanical item analysis procedures; but the causes of these difficulties need to be thoughtfully analyzed in order for improved test questions to be generated.)

The Concern over Test Bias

Reflecting older complaints that IQ tests measure language maturity more than native intelligence, various critics are assailing the widespread bias in standardized educational tests. Houts, for example, marshals a formidable array of educators and researchers who assert that present-day educational tests exacerbate social inequities by stigmatizing children instead of reducing such differences by encouraging a modification of the curriculum to "meet variations in interests, talents, backgrounds" (Houts, 1977; Leach, 1979). In their carefully documented book, Oller and Perkins (1978) together with

other researchers note the content similarity in achievement batteries, intelligence tests, personality inventories, and language proficiency tests. Then, through statistical analysis, they demonstrate that language proficiency is a major variable in the tests evaluated (accounting for .58 to .88 of the variance in the subscores of the California Achievement Test, for example). These writers conclude that, even for native speakers, tests that purportedly measure intelligence, or achievement, or personality, may primarily be measuring language proficiency. Nonnative speakers are therefore in real danger of being improperly evaluated by such instruments:

> It is obvious that the student who speaks and understands the language variety of the test will have an advantage over the student who is more familiar with a different variety. That is, the tests are clearly biased against speakers of nonmajority varieties of English. (Oller & Perkins, 1978, p. 34)

So great has the concern become over educational inequities that legal actions have been taken to protect against biased testing. In 1972, the Supreme Court focused on testing that places students in classes for the educable mentally retarded (EMR). The Court determined that the tests were culturally and socio-economically biased, and it ordered that the students be evaluated by unbiased instruments. When retested with culturally unbiased instruments, two-thirds of the Black students tested out of EMR into regular classes (Nakano, 1977, pp. 221-222). Also recognizing the limited language proficiency of many immigrants and minority groups, Title VII of the Civil Rights Act of 1964 "mandated employers to cease the use of test results to discriminate among potential employees if and when job relatedness could not be established" (Garcia, 1979a). In recent years, court cases and legislation focusing on tests with language bias have increased. These actions are intended to eliminate various educational inequities, freeing children of faulty and stigmatizing labels (for example, "mentally retarded") or halting the practice of educational "streaming" into inappropriate programs. But the broader purpose of these legal actions related to testing is to preserve individual rights not only in the area of citizenship and voting but also in the realms of education, employment and human dignity (Garcia, 1979b).

It has also been charged that certain reading comprehension tests evaluate cultural knowledge as well as language proficiency and "thereby discriminate against ESL students"; several tests were shown to contain 10 to 15% biased items (Mohan, 1979, p. 173). Consider such items as the following:

"There are red and white stripes and white stars in our flag. Our flag contains one . . . for every state.

(a) stripe (b) star . . ."

"Pam went to the party with a tall pointed black hat, long cape and a broom. She was dressed as a . . .

(a) witch (b) ghost (c) cowgirl (d) pumpkin"

"Bill ran out on his front porch to watch the firetruck. He lives in . . .

(a) a big apartment (b) a city house (c) a trailer . . ."

(Mohan, 1979, pp. 171, 172)

A form of cultural bias resulting from test administration procedures is provided by Briere in his evaluation of the Wepman Auditory Discrimination Test:

> As part Mohawk Indian, I have been trained never to trust a stranger behind my back in a small room. This sort of thing is frequently true of American Indians, Blacks, and Chicanos. Since the test administrator (tall, big, authoritative, and threatening) stands *behind* the children during the test (presumably to avoid lip cues), I wonder how valid the results are with groups who have been culturally trained *not* to allow a threatening figure to get behind them? (1973, p. 219; see also Labov, 1972, pp. 59-67)

It has even been suggested that test bias may be found in the very form that language tests take, specifically in integrative and discrete-point formats. Farhady (1979a) points to evidence indicating that all-discrete-point or all-integrative tests seem to discriminate against students of certain nationalities. For example, in one study, students from Israel and France scored significantly higher on an integrative placement test than on a parallel discrete-point test. But students from Taiwan and Korea scored significantly higher on the discrete-point form. Farhady also suggests the possibility of a sex bias related to test form (1979a; see also Farhady 1979b, pp. 162-170).

Still another concern is that of eliminating random bias related to the administration and scoring of language tests. For instance, oral test examiners who speak the native language of an examinee sometimes

subconsciously overlook certain errors simply because they encounter these so frequently in the classroom, while errors made by others may be looked at more critically. Moreover, listening tests administered in large rooms where the sound reverberates can result in a weaker showing that might otherwise be the case in a room with good acoustics.

A final potential for bias is the impact of the exam on the individual. Last year at a testing conference in Israel, an unusual theme emerged which focussed on the unintended effects of language tests (Cohen, 1984, p. 23). Studies in recent years have looked at the effect of text anxiety on performance on various types of language tests (Jones, Madsen, & Brown, 1980; Scott & Madsen, 1983; Stevenson, 1979). In one study at an English Language Center, 146 ESL students were administered a special anxiety scale. Roughly one-third of these people, it turned out, were quite anxiety prone when taking tests. The weaker performance of the anxiety-prone students on the most stressful subtest demonstrated that very stressful subtests constitute a source of hidden bias on language exams (Madsen, 1982, pp. 133-143).

Use of Rasch Analysis in Identifying Item Bias

We have seen that test bias can take many forms, ranging from items with built-in cultural assumptions to questionable administration procedures. Rasch software is particularly effective at identifying bias at the item level: not only items inappropriate for a given student population but also individual students whose background is not appropriate for the items one is using.

To demonstrate how Rasch is utilized in evaluating bias, we will first provide a brief introduction to the software program. We will conclude with a presentation of the study itself.

Microcomputer Rasch Calculations

Until very recently, calculations involving the Rasch one-parameter logistic model were done either by utilizing hand calculators or by means of main-frame computers with programs such as BICAL (Wright & Mead, 1976). Now microcomputer software has been produced to perform Rasch analysis, which promises to expand greatly the applications of latent trait measurement. Utilized in this study is Microscale Plus™, a microcomputer software program written for the IBM-PC. It was developed by Mediax Technologies, Inc. and Benjamin Wright, the author of BICAL, to perform the Rasch analysis. Although the program is presently somewhat limited as to the amount of data it can process at one time (254 students and 62 items, or vice versa), it is being upgraded to handle up to 200 items administered to an unlimited

number of students, depending on the user's disk storage capacity.

Data are entered into a spreadsheet program (Supercalc3™ currently) either manually using the keyboard or in comma-separated format from a mainframe computer. The spreadsheeet allows the user to arrange and correct data, specify headings and subheadings, and perform some mathematical calculations on the data. Each response is input into the data grid as either a 1 (correct response) or a 0 (incorrect response). Unlike many statistical analysis programs, Microscale will accommodate up to 50% missing data, estimating a student's score based on his performance on the rest of the test.

After the data have been input, they are analyzed by the computer to obtain a linear comparison of student abilities and item difficulties. The results of the analysis are displayed in several formats. This information can be presented on the computer screen or printed as hard copy using a dot matrix printer or a plotter. A tabular representation of the data is given in results tables as shown in Figure 1.

In the case of the item results table, the first column of the table lists the test item number. The second column shows how many students answered that item correctly, and the third column gives the estimated difficulty of the item in logits.[1] The standard error of the difficulty estimate is given in the fourth column. The "fit" statistics, which indicate how well each item is able to measure the individuals taking the test, are presented in the next two columns. These figures are based on all persons completing the specific item. The outfit statistic is significantly affected by unexpected responses, or "misfits." Outfit item plots emphasize questions that are answered by students whose ability level does not match closely the item's difficulty level. Such behavior as carelessness and lucky guesses are reflected in high positive outfits.

The infit statistic, on the other hand, is an information-weighted fit statistic emphasizing how well an item performs for students whose ability level corresponds closely to the difficulty of the item. Large positive infit values are associated with items that fit the model poorly. The statistics in the student results table are similar in nature to the item statistics but are based on student performance.

In addition to the results tables, graphic representations are available in which fundamental relationships in the data can be visualized; these plots allow one to quickly identify problem areas. Figure 1 is an outfit by logit plot for items, while Figure 2 shows outfit by logit for students on an ESL grammar test. Slightly resembling a scatterplot, these plots, when tightly clustered, indicated that the majority of questions are suitable for linking routines or for use in evaluation measures, or that the students are behaving as expected on the test. Information about goodness of fit can be useful, among other things, for

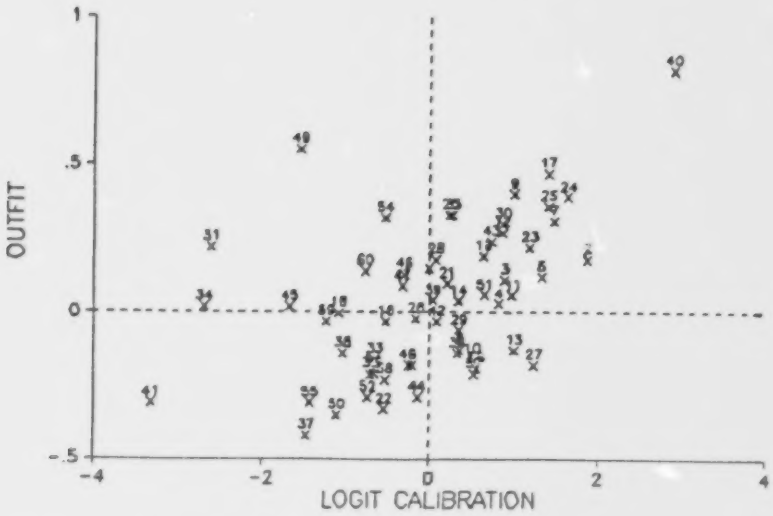


Figure 1. ESL Low Battery Grammar Test Outfit by Logit for Items

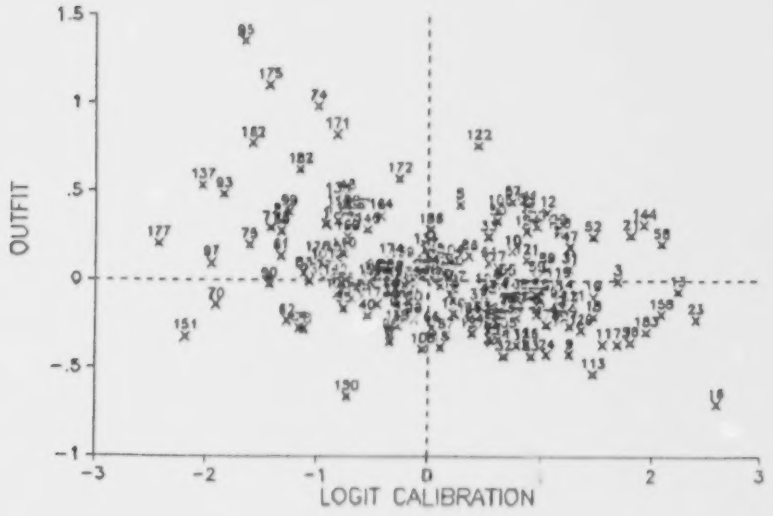


Figure 2. ESL Low Battery Grammar Test Outfit by Logit for Students

identifying items that are biased toward certain groups of students or for isolating students for which the test is not appropriate.

In addition to the outfit and infit plots, Microscale produces a "map" that juxtaposes items and students along the same scale (see Figure 3). This provides a clear, visual comparison of how well the test measures the full ability range of the students who took that test. A tightly-spaced rectangular set of items (such as when the lower plotting looks like a histogram that is flat), extending beyond high and low students, indicates that the test provides ideal coverage of the students being assessed. By studying the map, the tester is able to identify items that are of little or no use or to determine areas where additional items are necessary in order to have a complete and accurate assessment of the students' abilities. This information is extremely useful in expanding, refining, and revising the test.

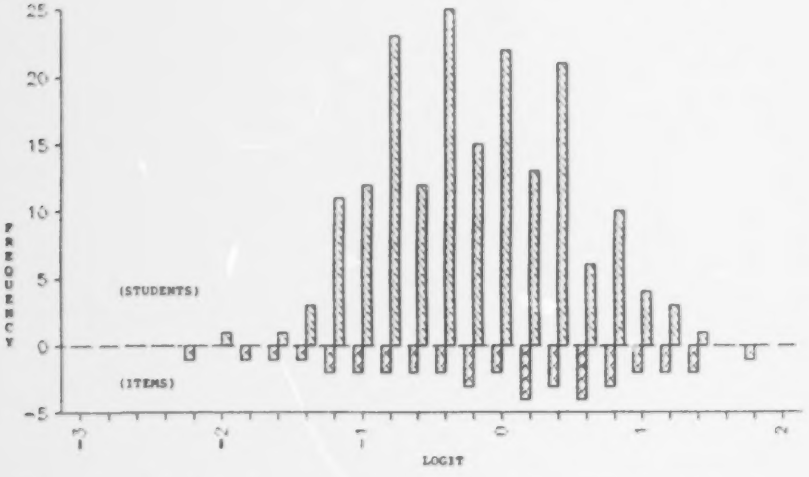


Figure 3. ESL Low Battery Listening Test Map of Student Ability Compared to Item Difficulty

The Bias Study

The bias study, utilizing Rasch analysis by means of the Microscale Plus program, consists of three procedures: first, a short contrived-bias check, then a simulated-bias computer run, and finally the principal investigation. the first two were planned simply to verify Microscale's sensitivity in identifying known bias.

The Contrived Bias Check

Since the administration of a biased exam would not be defensible, it was arranged to administer a Spanish-cognate vocabulary exercise in the reading labs at our English Language Center, where performance would not be counted toward student grades and where the items could be used for vocabulary building work.

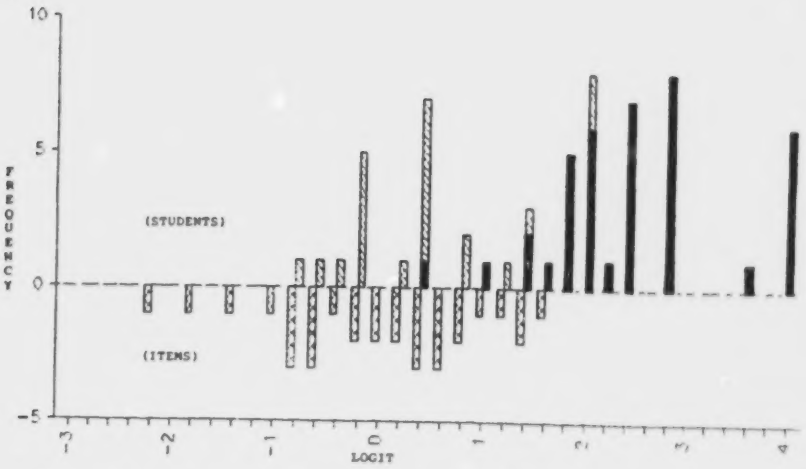
The 60 ESL students in the bias check included 39 persons who would presumably recognize the Spanish-cognate items because of their language background (36 native Spanish speakers, 2 Portuguese speakers, 1 Quechua speaker with considerable proficiency in Spanish); and 21 persons who would presumably not easily identify the items (10 Japanese speakers, 4 Chinese, 2 Arabic, 1 Farsi, 1 Korean, 1 Vietnamese, 1 Thai, 1 Greek). These 60 persons were drawn from 6 levels of instruction – Spanish speakers predominating in number at each level.

The 30-item, multiple-choice Spanish cognate exercise provided four options with a single-word stem. The test was presented in English and also in versions utilizing native-language instructions and options in Spanish and Japanese (there being only Spanish and Japanese students on the beginning level). The mean for the Spanish-Portuguese group was 26.4; the mean for the non-Spanish group was 17. For all the students combined, the mean was 23.1 (77%); the standard deviation was 5.7, and the reliability .86.

A Rasch analysis of the data did pick up the bias: first, in the map, which predictably generated a bimodal distribution of students (Figure 4). Then in the outfit display, a majority (73%) of the outliers, or persons who did not perform as expected statistically, proved to be non-Spanish speakers, even though this group constituted a mere third of the subjects evaluated (Figure 5). While less than half of the non-Spanish/Portuguese speakers were picked up by this routine (a few advanced non-Spanish speakers scoring in the 80's), the results are still rather dramatic: the bias is unmistakably identified.

The Simulated Bias Check

Our first simulated bias check consisted of performing a Rasch analysis on a 60-item, multiple-choice ESL grammar test. We then selected an easy item (No. 11) which had "behaved" normally, and we "*caused*" five students from the top quartile to miss this easy item by modifying the data as if they had answered incorrectly. (Let's assume these students were five non-science majors in this group of over 150 students and that item 11 was an easy scientific term.) The Rasch analysis that was performed this time now picked up item 11 as an outlier (Figure 6). In a real-life situation, a survey of which students missed the item could be effective in disclosing bias.



NOTE: The solid lines represent students with Spanish-speaking background.

Figure 4. Spanish Cognate Vocabulary Exercise Map of Student Ability Compared to Item Difficulty

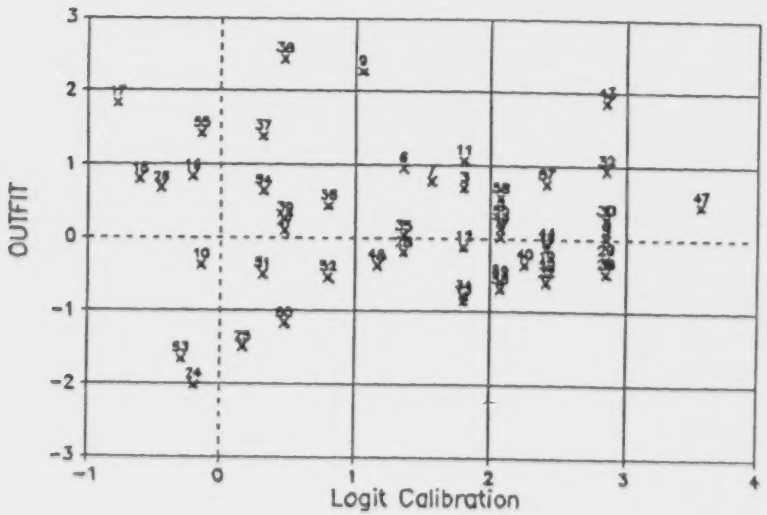


Figure 5. Spanish Cognate Vocabulary Exercise Outfit by Logit for Students

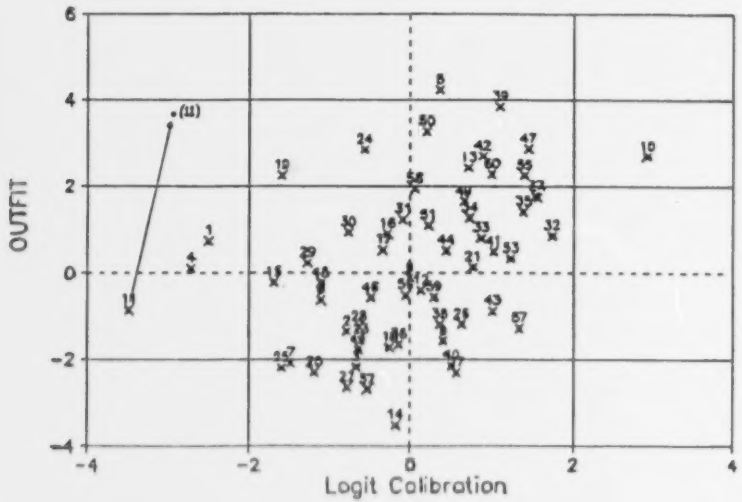


Figure 6. ESL Grammar Test Outfit by Logit for Items

Using the same set of data, we examined this time the *student* outfit printout and identified two persons who were in the lowest quartile of ability – students 23 and 32. Manipulating their test scores, we arranged to have both students get five difficult items right. Rerunning Microscale, we found their outfit scores jumping from approximately 1 to substantially above 4 – a very dramatic increase (Figure 7). This clearly flags both students as performing quite unexpectedly and, thus, facilitates the identification of cheating or bias.

The Principal Investigation

Having examined simulated bias and contrived bias, and having found the analysis procedure to be accurate, we turn now to an examination of real bias in ESL tests. It was predicted that there are identifiable patterns of poor fit statistics, and that one such pattern is associated with student language background. It was hypothesized as well that there are patterns of poor fit statistics that transcend language background: i.e., academic major, ability level, etc.

ESL students at three levels of proficiency, ranging from beginning to intermediate, were administered a five-part ESL placement battery. The three multiple-choice sections of this test were Rasch analyzed to identify possible bias, particularly in terms of language background. Analysis of variance was also utilized to confirm observed differences.

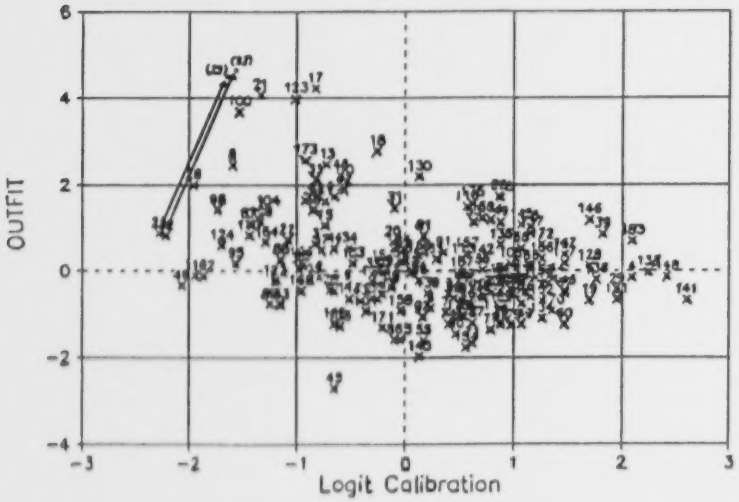


Figure 7. ESL Grammar Test Outfit by Logit for Students

Subjects

The 183 ESL students taking the tests ranged in ability from beginners to intermediate: 30% (54 students) were beginners – at Level 1; 40% (74 students) were lower intermediate – Level 2; 30% (55 students) were intermediate – Level 3. Of the 183, 101 or 55% were Spanish (56 males and 45 females, the average age being 20.4); 49 or 27% were Japanese (14 males and 35 females, the average age being 24.3); and 33 or 18% were native speakers of other languages (Portuguese, Korean, Chinese, Thai, French, German, Polish, Cambodian, Vietnamese, Indonesian, Arabic, Persian, Greek). Of these 33 students, 18 were male and 15 female, the average age being 24.9.

Measurement

Students were administered a 5-part placement battery: oral interview, dictation, grammar, listening, and reading; the latter three multiple-choice sections being analyzed as part of this study. The three multiple-choice subtests consist of the following: a 60-item, 25-minute grammar test consisting of 4-option sentence-completion questions; a listening test consisting of 20 3-option appropriate response questions and 20 4-option appropriate paraphrase items; a 60-item, 40-minute reading test consisting of 3-option paraphrases of highly embedded sentences.

Data Analysis

To avoid violating the unidimensionality assumption, the authors Rasch-analyzed the three subtests separately. Of the various data generated, we focused on the outfit statistics, as they are most sensitive to erratic, unpredicted behavior. Moreover, outfit statistics were used as the dependent variable in analysis of variance runs, with language background as the independent variable and level as moderator variable.

Results and Discussion

For the reading test, the mean was 32.7 (55%), the standard deviation 8.47, and the reliability .81. The listening exam generated a mean of 17.65 (44%), a standard deviation of 5.45, and a reliability of only .69. For the grammar test, the mean was 29.85 (50%), the standard deviation 11.63, and the reliability .90.

It was anticipated that outfit plottings of Rasch data by language groups would show a disproportionate number of outliers according to language background. This, however, was not the case, and expected differences were not substantiated (Figures 8, 9, 10). ANOVA calculations on outfit data by language groups likewise failed to indicate

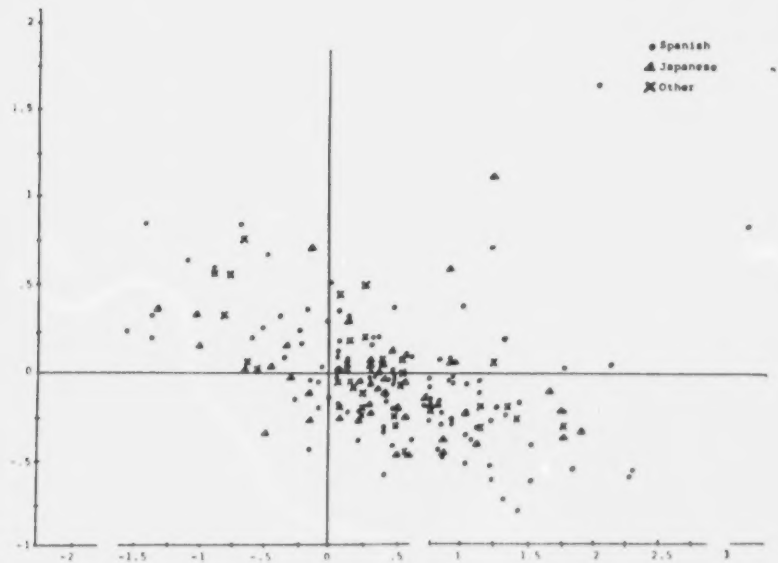


Figure 8. ESL Reading Test Student Distribution by Language

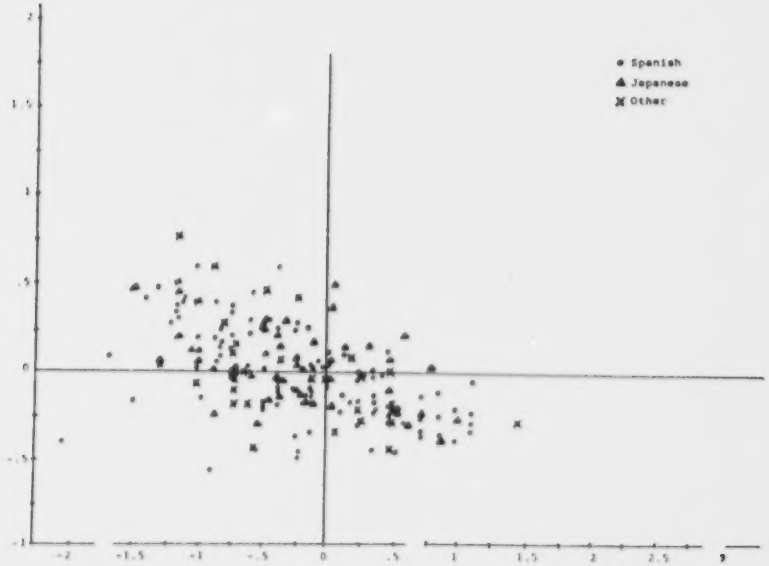


Figure 9. ESL Listening Test Student Distribution by Language

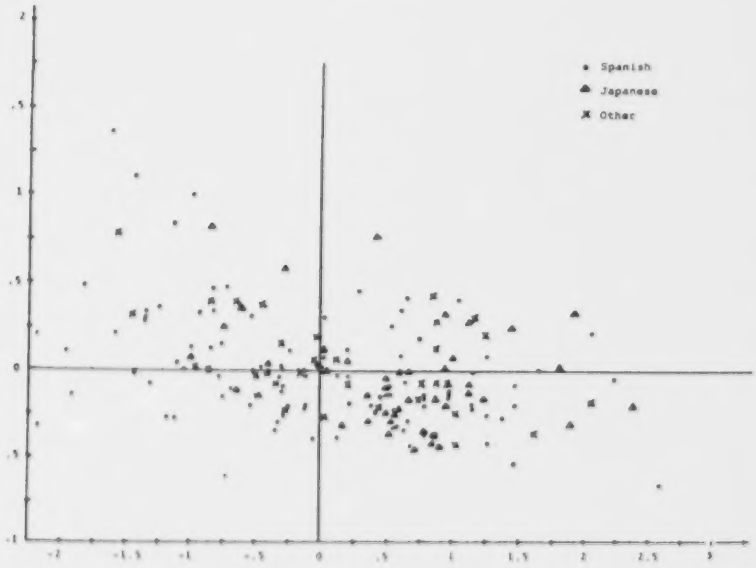


Figure 10. ESL Grammar Test Student Distribution by Language

statistically significant differences. Consequently, the initial hypothesis of differences in fit among language groups was not confirmed.

Turning to Rasch analyses of the three subtests, we next looked for patterns of outliers in areas other than language. On the reading test, outliers were distributed somewhat evenly from low ability students, at the left, to high ability students, at the right (Figure 11). But outfit data on the listening test patterned differently: here, outliers emerged almost exclusively from the lower half of the students (Figure 12). And on the grammar test, outliers were predominantly from the lowest third of the students (Figure 13). The ANOVA calculation, with proficiency level as independent variable and outfit as dependent, tended to corroborate the Microscale Plus displays of outliers: statistically significant fit differences occur at varying ability levels (Figure 2).

The greater amount of erratic behavior by low-ability students suggests that the grammar and listening subtests (the former in particular) may be too difficult for a number of beginning-level people. And in this sense these particular tests appear to be somewhat biased in favor of intermediate language learners. This finding tends to support the second hypothesis of outlier patterns not related to language background.

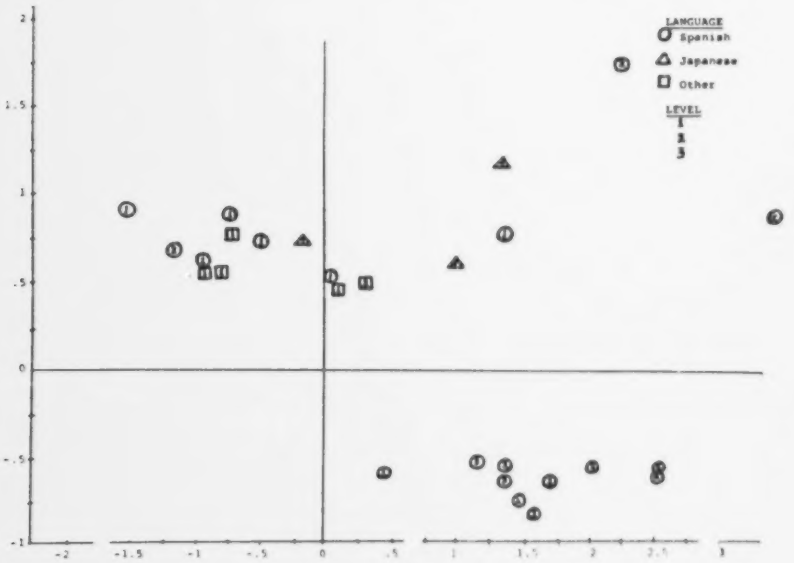


Figure 11. ESL Reading Test: Student Outliers by Language and Level

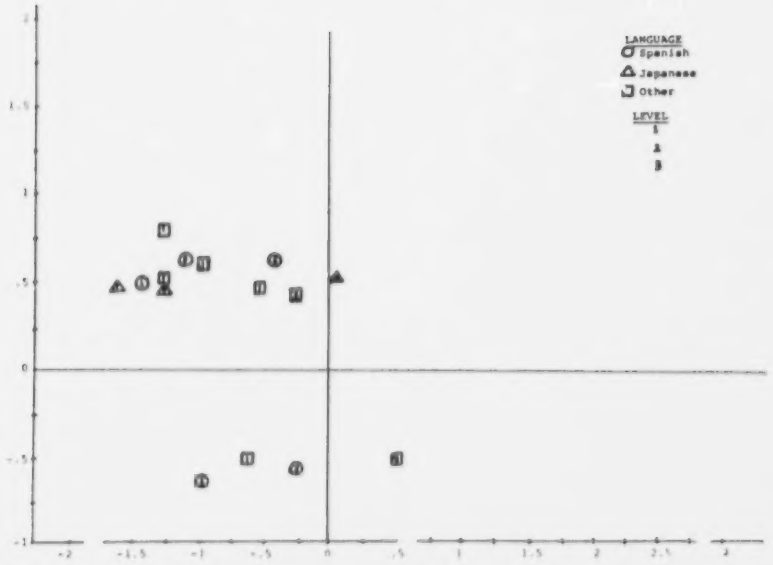


Figure 12. ESL Listening Test: Student Outliers by Language and Level

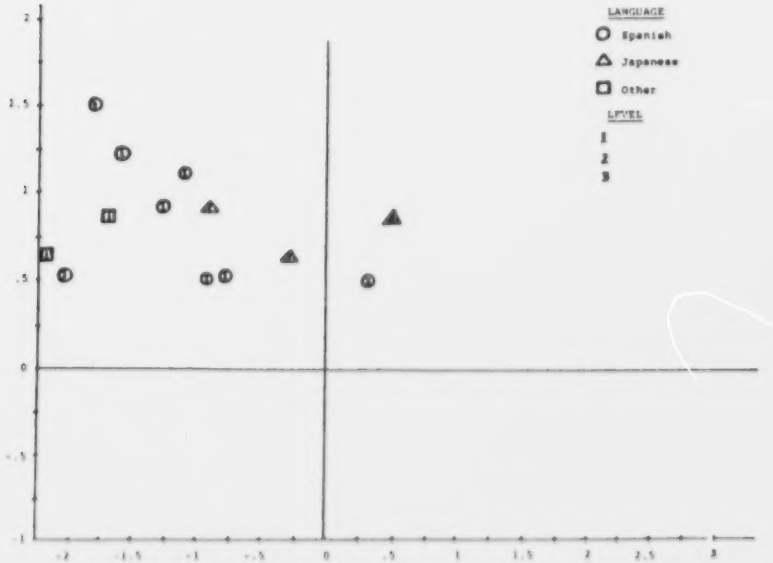


Figure 13. ESL Grammar Test: Student Outliers by Language and Level

It should be acknowledged, of course, that the findings in this study are rather test specific. The fact that the objective sections of this test battery appear to be relatively free of bias in favor of a particular language group (hypothesis 1) obviously does not relieve us of the need to scrutinize other ESL or foreign language exams for possible bias related to the language background of students. And the pattern of outliers found in connection with the subtests cannot be generalized to other ESL tests. Each exam can profit from such scrutiny. Moreover, a seemingly bias free exam that has been administered to one population could well contain bias when administered to another population.

In addition to *patterns* of bias in second language tests, of course, there are important individual instances of bias that the Rasch procedure is able to identify. The data presented here on simulated and contrived bias as well as that generated in virtually every outfit display show the potential of the one-parameter latent trait (Rasch) model in flagging single items or individual students manifesting potential bias. Fortunately, the computer display that signals aberrant behavior of items or individuals – the outfit chart – is simply a routine segment of Microscale's Rasch analysis that facilitates, among other things, accurate calibration of items and students, item banking, and even computerized adaptive testing.

But as helpful as the Microscale routine is, the teacher's role in discovering bias is vital. The outfit display does not identify the cause of erratic performance; this might be due to carelessness, cheating, guessing, unusual instructional background, etc. as well as to test bias. Where patterns of misfitting students (or items) occur, as in the case of the low-proficiency subjects in this study, the cause of aberrant performance may be relatively easy to identify. But uncovering bias in other instances may entail re-evaluating specific test items, looking at student background, and even interviewing certain examinees. While neither a routine nor a simple task, the pursuit of bias most assuredly contributes to the pursuit of excellence in language test construction.

References

Briere, E. J. (1973). Cross-cultural biases in language testing. In J. W. Oller, Jr. & J. C. Richards (Eds.), *Focus on the learner: pragmatic perspectives for the language teacher* (pp. 214-228). Rowley, MA: Newbury House.

Cohen, A. D. (1984). Fourth ACROLT meeting on language testing. *TESOL Quarterly, 18*(2), 23.

Englehard, G., Jr. (1980). An introduction to Rasch measurement and its application to test equating in the comprehensive assessment program. ERIC Document 194 636.

Farhady, H. (1979a). The disjunctive fallacy between discrete-point and integrative tests. *TESOL Quarterly, 13*(3), 347-357.

Farhady, H. (1979b). Test bias in language placement examinations. In C. A. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79* (pp. 162-170). Washington, DC: Teachers of English to Speakers of Other Languages.

Garcia, G. N. (1979a, March). Testing in the second language: mandates and decrees. Paper presented at the 13th Annual TESOL Convention, Boston.

Garcia, G. N. (1979b, July). Three wrongs can make a wrong. Address delivered at the TESOL Summer Institute, University of California at Los Angeles.

Hambleton, R. K. & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14*(2), 75-96.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing, 1*(2), 123-133.

Houts, P. L. (Ed.). (1977). *The myth of measurability*. New York: Hart.

Jones, R. L., Madsen, H. S., & Brown, B. L. (1980, March). *Evaluating affective variables in second-language testing*. Paper presented at the 14th Annual TESOL Convention, San Francisco.

Labov, W. (1972, June). Academic ignorance and black intelligence. *Atlantic Monthly*, pp. 59-67.

Leach, J. N. (1979, March). *Bias in standardized testing: an update*. Paper presented at the 13th Annual TESOL Convention, Boston.

Madsen, H. S. (1982). Determining the debilitative impact of test anxiety. *Language Learning, 32*(1), 133-143.

Mohan, B. (1979). Cultural bias in reading comprehension tests. In C. A. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79*, (pp. 171-177). Washington, DC: Teachers of English to Speakers of Other Languages.

Nakano, P. J. (1977). Educational implications of the Lau v. Nichols decision. In M. K. Burt, H. C. Dulay, & M. Finocchiaro (Eds.), *Viewpoints on English as a second language*, (pp. 219-234). New York: Regents Publishing.

Oller, J. W., Jr. & Perkins, K. (1978). *Language in education: testing the tests*. Rowley, MA: Newbury House.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Rubin, L. S. & Mott, D. E. W. (1983). Comparison of three different item banking methods for longitudinal test equating. (April) ERIC Document 228 315.

Scott, M. L. & Madsen, H. S. (1983). The influence of retesting on test affect. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 270-279). Rowley, MA: Newbury House.

Stevenson, D. G. (1979). *The experimental evaluation of test affect*. Unpublished master's thesis. Brigham Young University.

Wright, B. D. & Mead, R. J. (1976). BICAL: calibrating items with the Rasch model (University of Chicago, Research Memorandum No. 23). Chicago Department of Education, Statistical Laboratory.

Footnotes

[1]Logit is a unit of measure used by Microscale Plus. For further explanation, see Wright, B. D., & Stone, M. H. (1979). *Best test design* (pp. 16-17). Chicago: Mesa.

Item Banking Via DBASE II: The UCLA ESL Proficiency Examination Experience

Grant Henning
University of California

Abstract

In recent years a lot of interest has been focussed on how best to develop and maintain test item banks for use with individual ESL programs. The merging of developments in microprocessing technology with developments in measurement theory concerning latent trait measurement has enabled programs to maintain such item banks and to tailor tests to the needs of existing programs.

This paper reports on procedures developed at UCLA for item banking and machine construction of language proficiency tests using DBASE II software and microprocessor hardware. These procedures include the storage of Rasch model difficulty calibrations, a variety of other item statistics, content specifications, and item deployment data.

The advantages of this approach are:

1. Use is made of widely available software.
2. A variety of microprocessors can be used.
3. Large databases of more than 65,000 items may be stored, manipulated, and retrieved.
4. Alternate forms of tests can be constructed easily.
5. The tailoring of tests to different groups is facilitated.

The information presented is based on development efforts with the English as a Second Language Placement Exam (ESLPE) in use at the University of California at Los Angeles. Discussion centers on specific problems to be overcome in the organization and maintenance of such an item bank. The appropriateness of the measurement model employed for item bank construction and maintenance is discussed.

Item Banking Via DBASE II: The UCLA ESL Proficiency Examination Experience

The majority of moderate to large scale language teaching programs satisfy the need for proficiency/placement testing in one of three ways: the single reusable test approach, the multiple alternate forms approach, or the item banking approach. It should be understood from the outset that item banking as an approach to testing is far more dynamic than merely the storing of a particular group of test items for retrieval at a later time. Item banks are expandable aggregations of test items that have been calibrated and linked to a common scale or latent continuum, in much the same way that financial institutions gather and manipulate wealth in a common currency.

The Single Reusable Test Approach

Many programs still depend on a particular test existing in a single form in order to make proficiency assessments and, subsequently, placement decisions about students. Even if the test used exhibits a high degree of reliability and validity for these purposes, serious problems with this approach are inevitable. Repeated use increases the threat to test security as test papers may be lost or copied. Students who repeat the same test may obtain artificially high scores due to practice effects and test familiarity. Finally, as test content becomes generally known, the test will sooner or later become invalid.

The Multiple Alternate Forms Approach

According to this common approach, a variety of forms of the test are developed or obtained to satisfy the measurement needs of the program. Since these forms are used interchangeably, their critical, common characteristic is that an examinee is expected to obtain the same score regardless of which form of the test is given. Thus, the test forms either must have the same difficulty and range or be equated by means of a common scoring table.

Alternate forms are generally developed according to one of the following five methods:

1. The option sequence method. Multiple-choice options are randomly resequenced for each item on the various forms of the test.

2. The item sequence method. The various forms of the test differ only in the sequencing of the items.

3. The item pool method. Items are developed according to rigid specifications for the measurement of each desired objective. The items so developed form item pools. Alternate forms are developed by randomly sampling the same numbers of such items from the same pools.

4. The equated forms method. Two or more similar tests are developed and are administered to the same large sample of individuals from the intended test population. A conversion equating tale is derived to enable conversion of the obtained raw scores to a uniformly interpreted scale.

5. The equivalent forms method. Two or more similar tests are developed and administered to the same large sample of persons. The tests are altered and readministered until they satisfy the stringent criteria of equal means, equal variances, and equal covariances.

It is immediately obvious that the first two methods, involving the mere resequencing of items and options, do not alter the content of the test. Therefore, while these methods may help prevent collaboration during testing, they fail to minimize security breakdown, or familiarity and practice effects. Furthermore, there is evidence that the resequencing of item options frequently alters item difficulty and, to some extent, test equivalency.

The item pool method has been shown to produce items and forms of varying difficulty, in spite of efforts to prepare items according to the same specifications (Williams & Slawski, 1980). The equated forms method is well known to be costly and time-consuming because of its requirement that large samples be administered to all of the items on all forms in order for the forms to be equated. Also, once the equating has been completed, it is often impossible to generalize to persons beyond the equating sample, or to combinations of items other than the combination unique to the equating study. Finally, the equivalent forms method introduces criteria that are so rigid that in practice they can almost never be satisfied. These deficiencies suggest that serious consideration be given to item banking as an alternative approach.

The Item Banking Approach

This approach makes two critical assumptions necessary as conditions for implementation. First, item banking assumes that a theoretical procedure exists for positioning all present and future items in the bank on the same latent ability continuum. All items in the bank

become mathematically related as they are calibrated on the same latent trait. Secondly, item banking assumes that appropriate computer hardware and software is available for storing and retrieving all of the necessary information about the items.

Item banking, as it is considered here, satisfies these conditions in the ESL proficiency/placement testing program of the University of California at Los Angeles through the use of the Rasch one-parameter logistic model of latent trait measurement and through the use of widely available DBASE II software compatible with a variety of microcomputers.

Advantages of Item Banking

At UCLA, item banking has been found to have important advantages over the other approaches to proficiency placement testing considered earlier in this paper.

Because the Rasch Model allows for person-ability measurement that is independent of a particular test or combination of items, the item bank approach permits a nearly infinite variety of combinations and permutations of items to be drawn for any specific measurement purpose. This effectively minimizes security problems, since even those examinees with prior exposure to certain items in the bank cannot know what future combinations will appear.

The ability to draw items from the bank in any combination or permutation and to know in advance, on the basis of prior analysis, the degree to which the item difficulties will be appropriate to expected person–abilities is a prerequisite for computer adaptive testing. From such a bank items can be drawn successively to match the ongoing performance of the examinee. A person who succeeds (or fails) with one item will be presented with a more (or less) difficult item in an iterative fashion until the actual ability of the examinee is located on the latent ability continuum. This makes testing more efficient since each examinee is presented only those items that are relevant to the ability estimate of the particular examinee.

Through item banking, fewer items are needed to estimate the ability of any examinee. This has the potential to reduce testing time and to increase the reliability of person-ability estimates.

If precalibrated items are drawn from the bank and administered sequentially along the ability continuum, the final ability estimate does not necessarily require the computation of a total score. Ability is the point on the latent continuum where regular success gives way to regular failure in responding to items.

The addition of new items to the bank, and their requisite calibration on the same established ability continuum, does not require the readministration of all the old items along with the new items to the

original sample of persons. Thus, use can be made of a small, stable group of linking items, or persons, which permit the joining of all new items to the appropriate position on the established latent ability continuum. In this way, the cost of equating all possible forms of a test can be greatly reduced.

Characteristics of the ESL Proficiency/Placement Examination (ESLPE) at UCLA

The ESLPE contains five multiple-choice type subtests of 30-items each, and one 20-minute composition/writing task. The five objective subtests include listening comprehension, reading comprehension, grammar appropriacy, vocabulary recognition, and writing error detection – all presented in a four-option, multiple-choice format to facilitate machine scoring of answer sheets. For each academic quarter a new form of the test is introduced consisting of approximately 100 preanalyzed items from the item bank and 50 experimental items. During this administration the experimental items are pretested for possible inclusion in the bank. Global internal-consistency reliability estimates for new test forms regularly exceed 0.95, and reliability estimates for the 30-item subtests regularly exceed 0.85. In the fall of 1984, a preliminary study based on the performance of 100 UCLA graduate students on the ESLPE and TOEFL revealed a correlation of 0.79 between the two tests. A higher correlation would be expected for the entire cross-section of examinees.

The examinees are comprised of approximately 2,000 adult students per year who enter or transfer into regular university courses or university extension courses. Approximately 300 to 900 such students are examined each academic quarter. The decision to employ Rasch Model as the latent-trait measurement model for this project was made partially in consideration of the expected sample sizes for any given test administration, since the Rasch Model is the logistic model which allows item banking with the smallest number of examinees.

Table 1 displays the item structure of the ESLPE along with the ratio of preanalyzed to experimental items for each form of the test.

Structure of the Item Bank

In the UCLA example, DBASE II software is used for the storage and retrieval of items calibrated into the item bank. DBASE II, and its successor DBASE III, are currently among the most popular software programs for database management available for use with a wide variety of microcomputers. DBASE II permits the storage of 16 fields of information on up to 65,000 items. Each field may contain up to 256 characters or numerical bits of information. For items of short to moderate length, it is possible to store an entire item in one field of the

Table 1
Item Structure of the UCLA ESL Placement Examination

Subtest	N of Items	Preanalyzed Items	Experimental Items
Listening	30	10 10	10
Reading	30	6 6 6	6 6
Grammar	30	20	10
Vocabulary	30	20	10
Writing Error Detection	30	5 5 5 5	5 5
Totals	150	98	52

file, leaving 15 fields free to record information about the item. In the UCLA example, the item itself is not currently stored in the bank, but the identification number of the item is stored. This preserves the possibility of maintaining longer items and passages while at the same time allowing more available space for information about the item.

Once the database file has been created and the data has been entered according to the specified file structure, it is possible to call up items according to any stored set of specifications. In our example it would be possible to call up all the grammar items appropriate for students in any designated ability range. Simultaneously, items can be limited to prespecified boundaries of standard error of measurement, fit validity, point biserial correlation, and content area specification. And, at the same time, bias statistics can be reported for the use of the item with students from any particular native language background. In short, a group of items can be assembled by machine to form a test with prespecified standards of length, difficulty, reliability and validity, even though the particular group of items may never have been assembled or administered in that particular combination before.

The current file structure of the UCLA ESLPE item bank is reported in Table 2.

The various field names are used to label classification fields containing important available information on each item. Eventually augmenting these fields with further information on language group bias and susceptibility to test-wiseness strategies is planned.

Table 2
File Structure of the ESLPE Item Bank

Field	Name	Type	Width
001	IDNO	C	005
002	FORM	C	040
003	PASSAGE	C	005
004	CONT	C	008
005	DIFCAL	N	006
006	SEM	N	004
007	TFIT	N	006
008	TFITB	N	006
009	DISC	N	005
010	PBIS	N	005
011	ERIM	N	004
012	CLASS	C	010

Currently, the item bank contains only about 280 items, since this project was begun only seven months ago. The bank is growing however, at the rate of about 50 items per quarter. The usefulness of the bank for a variety of applications should grow as its size increases. Eventually, the capability may exist to link ability estimates for students at a variety of institutions, a step that could greatly enhance the evaluation of participating language teaching programs. A preliminary step in this direction would be for interested institutions to develop similar item banks. These could be linked later to the same latent scale.

It should be noted that a variety of other software programs are available for item banking purposes (Deck & Estes, 1984). The MICROSCALE program developed under the direction of Benjamin Wright of the University of Chicago is one of the outstanding, though more expensive, packages available.

References

Arter, J. A. (1984, April). *Technical and practical issues in item banking*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Deck, D. & Estes, G. (1984, April). *Microcomputer software support for item banking*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Henning, G. (1984). Some advantages of latent trait measurement theory in the development of language tests. ***Language Testing***, *1*(2).

Williams, P. L. & Slawski, E. J. (1980, April). *Application of the Rasch Model for the development of equivalent forms of criterion referenced tests*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

Computerized Multilevel ESL Testing, A Rapid Screening Methodology

Marilyn M. Hicks
Educational Testing Service, Princeton, New Jersey

Abstract

The results of real data simulations indicated that computerized multilevel ESL testing could be an effective and relatively simple method of tailoring a language test to groups of individuals at varying levels of ability. Easily implemented, this testing procedure should be useful for screening and placement of large numbers of examinees. The utility of the procedure depends on an accurate equating methodology. IRT true score equating is probably the best equating method currently available for this purpose.

Computerized Multilevel ESL Testing, A Rapid Screening Methodology

Much of the recent computerized testing research has been directed toward tailoring tests for the individual examinee; however, a form of tailored testing in which tests are adapted to groups of individuals at various ability levels could be useful in certain situations. In this mode of testing, shorter versions of a given test of varying levels of difficulty are administered commensurate with the examinee's ability level. Implementation of this method of testing has been impeded by difficulties encountered in determining the appropriate level of the test to be administered (see Betz & Weiss, 1973; Linn, Rock & Cleary, 1969). Marco (1977) investigated multilevel testing, an adaptive strategy in which examinees were branched directly through a measurement test utilizing branching and stopping criteria that determined several levels of the test. In that study, a paper and pencil version of the test was administered in which examinees were required to branch themselves based on "guesstimates" of their performance which resulted in a high proportion of branching errors among low-ability students. A computer administered multilevel test, as simulated in this study, alleviated many of these branching errors, resulting in a simple and effective form of adaptive testing.

The testing procedure to be described might be successfully applied as a quick screening and placement strategy; consequently its major purpose is classificatory in terms of *levels* of performance as they relate to some full length measure.

Methods

Evaluation of multilevel testing in ESL was based on the responses of 1327 examinees (a spaced sample of the total domestic group of examinees) to 38 operational and 20 experimental items in Section II (Structure and Written Expression), and 58 operational and 30 experimental items in Section III (Vocabulary and Reading Comprehension) of the Test of English as a Foreign Language (TOEFL). Computer programs were written that simulated the administration of the levels of the tests. Tests were constructed using the items in the pool defined above, then each examinee record was checked for correct or incorrect responses, and the tests were scored accordingly.

Multilevel testing was simulated utilizing five delta (difficulty) levels as follows:

Delta ≤ 10.9	11-11.9	12-12.9	13-13.9	Delta ≥ 14.0
(A)	(B)	(C)	(D)	(E)

(Deltas are item difficulties that have been transformed to a scale with a mean of 13 and a standard deviation of 4.) For Section II, each delta level consisted of 5 items. Starting at level C, the examinee was branched up to the next unattempted level if 4 or more C-level items were answered correctly, and branched down to the next unattempted level for 3 or less correct responses. Testing was completed after 3 delta levels were administered resulting in three test levels: CBA, CDB or CBD, and CDE. Examinees were also presented 4 additional items from the levels not administered in order to facilitate the equating. For Section III, each delta level consisted of 8 items, and the examinee was branched up if 5 or more items were answered correctly, and down for 4 or fewer correct responses. As in Section II, 4 additional items were administered to improve the equatability of the levels. The rationale for these branching criteria was based on the differing scales of difficulty for the two sections. For a given scaled score, examinees must answer proportionately more items correctly in Section II than in Section III. Consequently, the branching criteria required more correct responses on Section II than on Section III, and thus approximating the scales of difficulty on the regular TOEFL.

Each test level consisted of 19 items for Section II, and 28 items for Section III, as compared with 38 operational items in Section II, and 58 operational items in Section III for the full length TOEFL. (Examinees are actually administered 40 items in Section II and 60 items in Section III on the regular TOEFL; however, 2 items in each section are not scored and are administered in such a way as to ensure the validity of the test scores.) The levels were equated to the TOEFL scale using IRT true score equating (Lord, 1980, chap. 13), consequently each examinee obtained two scaled scores, one for the full length TOEFL and another for one of the levels. The equating procedure was a necessary adjunct to the testing methodology, since the raw scores obtained on any of the levels are not interpretable in terms of ability. Utilizing true score IRT equating methods in which abilities rather than scores are equated, shorter tests can be easily scaled to longer measures, a process that is problematic with conventional equating methods.

The objective of the testing methodology simulated in this study was to branch examinees through item subsets of difficulty appropriate to their ability such that scores on the shorter tests were "comparable" to those on the full length TOEFL. Comparability was evaluated in terms

Table 1
Means, Standard Deviations and Correlations for Full Length and Multilevel TOEFL

		Section II		Section III	
		TOEFL 38 Items	Multi. 19 Items	TOEFL 58 Items	Multi. 28 Items
Total Group	MN.	49	49	49	49
	SD	7.7	7.9	7.4	7.2
	range	24-68	25-68	28-67	28-67
		r = .94		r = .95	
		N = 1327		N = 1327	
Level 1	MN	41	41	41	40
	SD	4.5	4.7	4.3	4.0
	range	24-56	25-52	28-52	28-48
		r = .84		r = .79	
		N = 410		N = 349	
Level 2	MN	49	49	49	49
	SD	4.5	4.0	4.3	4.2
	range	38-63	38-58	32-59	33-59
		r = .79		r = .83	
		N = 583		N = 634	
Level 3	MN	58	59	57	58
	SD	5.2	4.5	4.0	3.5
	range	43-68	48-68	45-67	50-67
		r = .85		r = .85	
		N = 334		N = 344	

of (a) the correlation of the two sets of scores, and (b) an evaluation of whether the examinee was branched to the appropriate test level using the full length TOEFL score as criterion.

Results

Summary data for the equated scores on the regular TOEFL and the test levels as given in Table 1 indicate that examinees were assigned to appropriate test levels with a high degree of accuracy. The within level correlations are attenuated since they are based on a restricted range of ability but do indicate the relative relationships between the levels and the full length TOEFL. Since the full length TOEFL is designed for

Table 2
Means, Standard Deviations and Correlations for Full Length and Multilevel TOEFL, Easier Branching Criteria

		Section II		Section III	
		TOEFL 38 Items	Multi. 19 Items	TOEFL 58 Items	Multi. 28 Items
Total Group	MN.	49	49	49	49
	SD	7.7	8.0	7.4	7.8
	range	24-68	25-68	28-67	28-67
		r = .94		r = .94	
		N = 1327		N = 1327	
Level 1	MN	39	38	39	37
	SD	4.2	4.6	4.1	3.5
	range	24-47	25-46	28-48	28-45
		r = .79		r = .73	
		N = 180		N = 176	
Level 2	MN	46	46	47	46
	SD	4.5	4.0	4.8	5.0
	range	35-59	32-56	29-59	28-57
		r = .80		r = .83	
		N = 568		N = 616	
Level 3	MN	55	56	55	56
	SD	5.8	5.2	4.9	4.4
	range	40-68	44-68	36-67	44-67
		r = .88		r = .88	
		N = 579		N = 535	

optimal measurement at the upper portion of the score scale, larger correlations are observed there. As indicated above, the criteria used to branch the examinees reflect the differences in the scales of the two sections on the full length TOEFL. The concurrence of the distributions of examinees on the two sections seemed to justify these choices.

The effects on the overall results of changing the criteria for branching were virtually nil, but the criteria did affect the distribution of examinees among the three levels. An easier set of branching criteria were analyzed in which examinees were branched up for 3 or more correct responses on Section II, and 4 or more correct responses on Section III, and branched down otherwise. The effect was to distribute more examinees in Levels 2 and 3 with a concomitant lowering of

means in all levels. As given in Table 2, these results are encouraging since they indicate that some examinees can be tested reliably on two adjacent levels of the test.

While a degree of spuriousness was introduced into these correlations by virtue of some item overlap, this does not seriously detract from the fact that an appropriate subset of items adequately reproduced the regular TOEFL score.

The delta levels of the three multilevel tests roughly correspond to the following scaled score ranges:

Test ABC	20 - 50
Test CDB	35 - 60
Test CDE	50 - 68

For the data in Table 1, 100% of the examinees were routed to the correct level based on these ranges, using scaled scores on the regular TOEFL as a criterion. Ninety-four percent of the scores on the multilevel tests fell within the 95% confidence interval for the standard error of measurement of the full length TOEFL score for Section II ($\pm$ 5.4) and 92% of the scores on Section III were within this confidence range ($\pm$ 4.6). All other scores were close to these boundaries.

Graphs of the equating curves for the test levels are presented in Figures 1 and 2. The utility of the equating procedure is evident from these graphs where any given number right score converts appropriately depending on the level of the test. For example, in Figure 2 a raw score of 15 yields a scaled score of 43 on Level 1, a scaled score of 48 on Level 2 and a scaled score of 52 on Level 3. The graph also suggests the nature of the "continuous" measurement inherent in this methodology; as fewer items are included in each of the delta levels and the number of delta levels increases, more test levels result. For most practical purposes, however, three test levels may prove to be sufficient.

Some Implications of the Results of Multilevel Testing

One implication of the outcomes of this simulation is the confirmation of the efficiency of tailoring tests to levels of ability. Fewer items are required if the item difficulties are adapted to ability level – the basic premise of adaptive testing. In adaptive measurement, this results in a more equitable distribution of the precision of measurement over the score scale. Information relative to the magnitude of the standard error of measurement over the three levels can be inferred from the relative efficiency curves given in Figure 3. (Relative efficiency curves are derived from the item data, and specify the relative

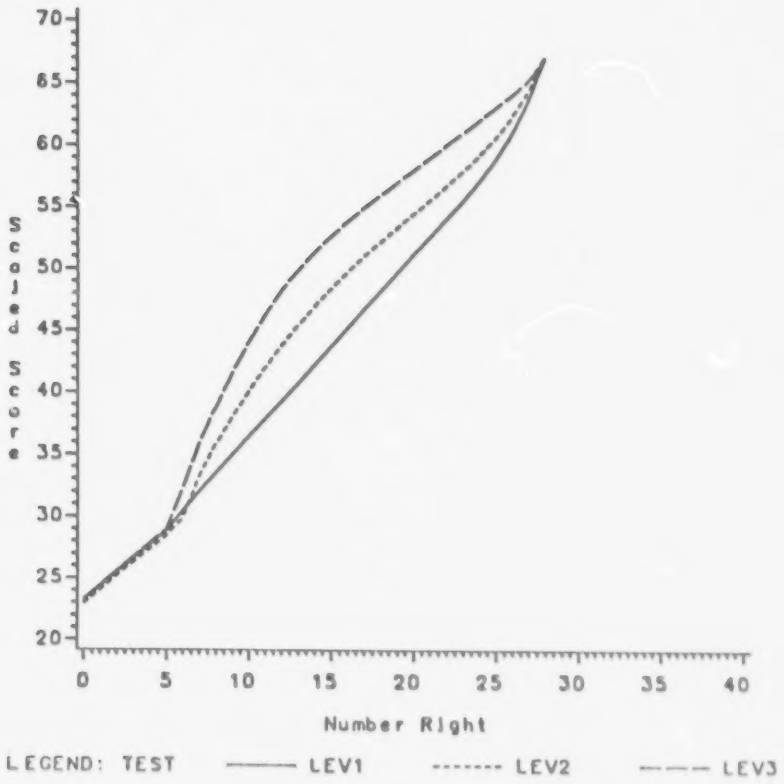


Figure 1. Equating curves for multilevel tests: Section III

effectiveness of measurement at any point on the score scale. See Lord, 1980, chap. 6). The chief reason for differences in the location of the maxima of these curves is the smallest error of measurement observed at those ranges, since relevant item parameter information is very similar across the levels. Relative efficiency curves depend mainly on discrimination and error of measurement, and discrimination parameters for TOEFL do not vary substantially over the range of difficulty. From the relationships among the curves, it can be seen that the precision of measurement has been equalized over the range of the scale. The maximum of a typical relative efficiency curve for the full length TOEFL falls somewhere between those observed for Levels 2 and 3.

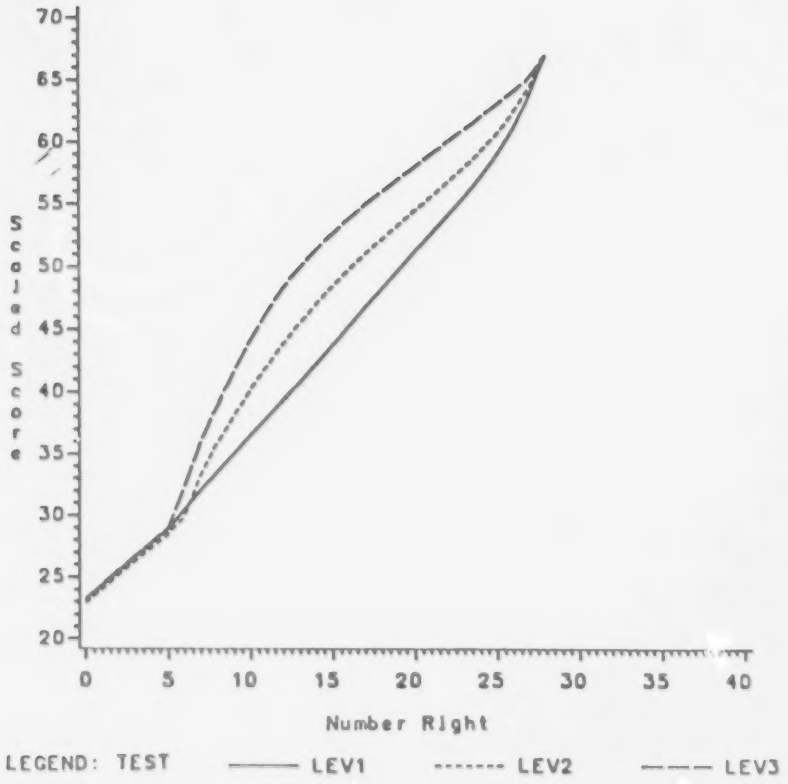


Figure 2. Equating curves for multilevel tests: Section III

In actual test construction, target relative efficiency curves would be specified and items selected to meet these specifications. In doing so, the tests can be constructed to produce a given rate of discrimination and a specified precision of measurement.

Each test level represented a sample of the content of the full length test. Nonetheless, this sample of approximately one half the regular test estimated performance on the latter rather well; that is, virtually the same information is provided by performance on a test with 38 items as on a test with 28 items since a total TOEFL section score and no other information about performance on the subsections is reported.

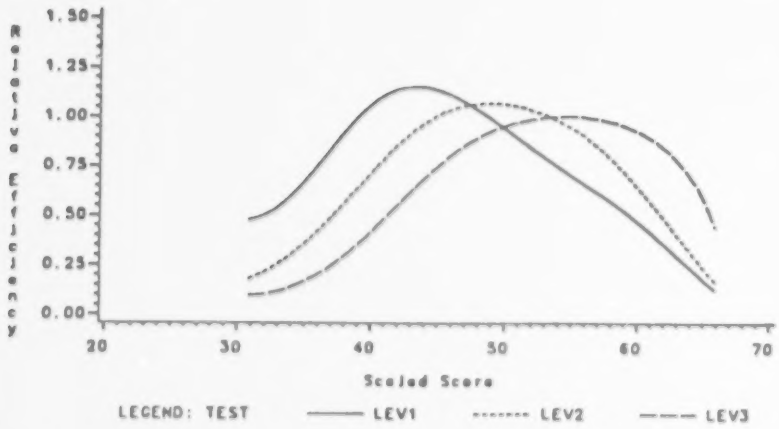


Figure 3. Relative efficiency curves for multilevel tests: Section II

Given that accurate and precise measurement can be obtained on shorter tests tailored for levels of ability, but that the shorter test must necessarily be only representative of the content of the longer tests, how can the scores be interpreted in terms of content of the full length test? It is probable that this could be easily accomplished by a re-examination of content areas in order to determine a hierarchy or overlap of content skills, and by a redefinition of the specifications in order to determine what broader categories of skills are represented. This will be necessary if shorter adaptive measures are to be interpretable or constructable in terms of the content of the longer test.

Discussion

The results of the simulations in this study demonstrated that a computer administered multilevel test can provide efficient and accurate measurement at differing ability levels. The correlations between scaled scores on the multilevel tests and the regular TOEFL (.94 - .95), and the close correspondence of the summary data indicated that this method of testing assigned examinees to appropriate levels with a high degree of accuracy. While the branching criteria affected the distribution of examinees among the levels, it did not impact on the overall results appreciably, demonstrating that the "continuous" test levels can provide adequate measurement at adjacent ability levels. Indeed, it easily handled examinees at the boundaries of the levels, the most difficult to classify in other methods of routing examinees to test levels.

The numbers of test levels can be easily increased to 4 or 5 by increasing the number of difficulty levels for branching. As a simple and effective method of adaptive measurement, computer administered multilevel testing should be particularly useful in the screening and placement of examinees into an appropriate language proficiency category.

References

Betz, N. E., & Weiss, D. A. (1973). *An empirical study of computer-administered two-stage ability testing.* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Linn, R. L., Rock, D. A. & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 29,* 129-146.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160.

Item Selection from Pretests in Mixed Ability Groups

John H. A. L. de Jong
CITO, Dutch National Institute for Educational Measurement
Arnhem, Netherlands

Abstract

Since 1983, two different types of secondary education in the Netherlands have been brought together in one type of school with a single curriculum. However, because of the span of the ability range, examinations at two different levels are offered at the end of the curriculum. Partial overlap of the examinations for each level, introduced for efficiency reasons, is justified by the overlap of the ability ranges at each level and permits equating procedures. Because the examinee's choice of level of assessment is based on his or her performance on parts of the examination, items are pretested on samples of about 300 subjects from the total mixed-ability group; but the items are evaluated as to their appropriateness for either level. The usefulness of item response theory (the Rasch model) for the selection of items on foreign language listening comprehension tests is demonstrated, and the appropriateness of the final versions of the tests for the levels targeted is evaluated.

Item Selection from Pretests in Mixed Ability Groups

In the Netherlands, the system of secondary education is highly differentiated. Four different types of schools produce four different foreign language proficiency levels.

LBO provides basic vocational training and is attended by about 22% of the age cohort. LBO is a four-year program. Final examinations are taken, as a rule, at age 16. The LBO foreign language program consists of about 300 hours of instruction. The final level is comparable to the "threshold level" as defined by the Council of Europe (Van Ek, 1975).

MAVO is the lowest level of general education. It is intended as a base for clerical jobs requiring further training and is attended by about 37% of the age cohort. MAVO is a four-year program. Final examinations are offered at two levels (see below) and are taken at age 16. The MAVO foreign language program consists of about 400 class hours. An average scaled score of 500 on TOEFL (Test of English as a Foreign Language, Educational Testing Service, Princeton, NJ) indicates the level of English proficiency of MAVO students who graduate at the higher level.

HAVO is an intermediate level of general education. It prepares students for higher nonacademic studies and is attended by about 22% of the age cohort. HAVO is a five-year program. Final examinations are taken at age 17. The HAVO foreign language program consists of about 500 classroom hours. Graduating students attain a proficiency level comparable to the higher levels of secondary education in most Western countries.

VWO is the highest level of general education. It prepares students for further academic study. It is attended by about 18% of the age cohort. VWO is a six-year program. Final examinations are taken at age 18. The VWO foreign language program consists of about 600 hours of instruction.

In the Netherlands, final examinations consist of an external part and an internal part. Examinations for the foreign languages cover four skills: reading comprehension, listening comprehension, writing, and speaking. In HAVO and VWO examinations, knowledge and understanding of literature are also assessed. Reading comprehension is assessed in the external exam, while the remaining skills are assessed by the schools. Most schools (about 95%) use the CITO tests (described below) to assess listening comprehension.

As stated above, examinations can be taken at two levels in the MAVO schools. Each student can graduate in each subject at two different levels: MAVO-C and MAVO-D. Since 1983, the MAVO-C level (which is identical to the LBO level) has been offered to all MAVO students to enable them to achieve a higher success rate. Though it is true that fewer students fail now than under the former system, the new system has a negative effect on the mean level of ability attained in MAVO schools (de Jong, 1985).

The curriculum in MAVO schools is identical for all students. During their final year, students receive two grades for all schoolwork as part of the internal school assessment: a MAVO-C and a MAVO-D grading. Before they sit for the external examinations, students must decide at which level they will take each subject. They base this decision on the school grades they have received during the year. A student who is good at listening comprehension may still decide to take the MAVO exam at the lower level because he or she stands no chance to pass at the higher level for any of the other skills. Conversely, a student who is not good at listening comprehension may decide to take the exam at the higher level because he or she expects to be able to compensate for this by getting high grades in the other skills. Therefore the distributions of students' ability in each of the skills at the MAVO-C and MAVO-D levels overlap to a large extent.

Examinations in the Dutch educational system are essentially norm-referenced, because generally accepted criteria have not been defined. In the MAVO schools, students decide at which level they will sit for this norm-referenced test based on the probability of success. The student's decision directly influences the evaluation, as the decision to partake in the examination at a certain level influences the distribution of students at that level. It is, therefore, essential that tests used in the schools be highly discriminatory at the expected cutoff scores of both levels.

The purpose of this paper is to illustrate the method used to determine the effectiveness of items at two different ability levels, given that these items are pretested on a sample of the total mixed-ability group and that the only information available is an estimated difference in mean ability and an estimated distribution of ability in the two target groups. After a description of the tests and the students, the underlying theory and the method used for item selection will be explained in nontechnical language. The paper concludes with an evaluation of the method as it was used in test design.

Tests

The tests used in this study are the regular CITO tests of foreign

language listening comprehension. The tests were originally developed in a research project at the University of Utrecht (Groot, 1975). Since Dutch examinations are public, new tests have to be developed each year. Tests of foreign language listening comprehension have been developed at CITO for use in final examinations in schools for over 10 years. The objective of the tests, as defined by Groot (1975), is to evaluate the learner's ability to understand the foreign language as spoken spontaneously by educated native speakers at normal conversational speed. Extremely informal elements, lexical and syntactical elements incomprehensible to less-educated native speakers, and topics requiring specific knowledge are excluded from the material in the tests. The material consists of three or four interviews with native speakers in various occupations or professions. The item format is a multiple-choice question with three options. The language material is divided into passages requiring 40 to 50 seconds to read; the correct answer is a one-phrase summary of the global contents of the passage. Each passage on the tape is followed by a pause of 20 seconds which allows the students to read the item in the test book and to mark the answer on the answer sheet.

A second item format, modified cloze with two options (de Jong, 1983, 1984), has been in use since 1981. In each 20 to 30 second passage, a word or group of words is deleted and replaced by an electronic sound. The words deleted from the original text were chosen for their semantic relevance to the context. The language material consists of discussions, conversations, and news or other radio programs. During a 7-second pause that follows each language sample, the students must decide which of the two options in the test book should be used to repair the text.

In 1983, a third item format was introduced especially for the evaluation of listening comprehension at the lower end of the ability continuum (Boeijen, 1984). Situations are presented in which nonnative speakers, speaking the target language at the threshold level (Van Ek, 1975), ask native speakers for certain services. The mastery of specific notions in the listening situation is measured by multiple-choice items consisting of two photographs, one of which represents the native speaker's answers.

Each test consists of at least three parts; thus, a number of different topics or fields of interest are covered. Depending on level and item format, test length varies from 40 to 50 items. Administration time of the total test varies accordingly between 50 and 60 minutes. Test reliability (KR 20) ranges from .75 to .85.

Tests are developed for the various levels in the Dutch educational system. The MAVO listening comprehension tests consist of five parts: A, B, C, D, and E. The results on parts A, B, and C determine the

marks at the MAVO-C level, whereas the results on parts B, D, and E determine the marks at the MAVO-D level. Thus, the two tests partly overlap, and part B figures in the test at both levels. This overlap was introduced primarily to reduce administration time.

Pretesting

Annually, each part of the new tests is pretested on separate samples of about 150 students from the target population. Whenever special attention is necessary, samples are doubled and grouped in such a way as to permit a missing data design. Special attention may be required by statisticians when tests using recently introduced item formats are pretested, or, as in the case of the MAVO tests, when tests are developed for a target population that cannot yet be traced.

The five parts of MAVO tests are pretested on separate samples of about 300 students who are in their final year in the MAVO schools. The samples are selected in such a way that each is composed of students from eight or more different schools. Pretesting takes place in the autumn, about four months before the administration of the tests in their final form. At that time which of the students in the sample will choose to take the exam at the C-level or the D-level is not yet known. However, some facts are known:

1. 18% to 20% will eventually take MAVO-C and 80% to 82% will take MAVO-D.
2. The difference in mean ability of the two levels is comparable to the distance between, for example, HAVO and VWO: approximately one unit on the logit scale, large enough to justify separate tests and small enough to allow equating procedures (cf. Petersen, Marco, & Stewart, 1982).
3. The difference between the ability of the students in the autumn and at the date of the final tests is abut one unit on the logit scale.
4. Because of the highly differentiated school system, the distribution of ability within each level is rather narrow (the standard deviation equals approximately one logit).
5. The distributions of ability of the two levels are bound to overlap, especially since the results on the listening comprehension tests only in part determine the level at which a student will eventually take the exam.

After the administration of the final version of the test, samples of 1,000 students are drawn to determine the cutoff score and the norms. Data from this sample are also used for evaluation purposes.

Method

The item selection method used is based on item response theory, using the Rasch model (Rasch, 1960). Test results were analyzed using the computer program CALFIT (Wright & Mead, 1975). Recalibration of items and test equating were performed according to the procedures described in Wright and Stone (1979). In the Rasch model, one item parameter and one person parameter determine the probability that a given examinee will answer a given item correctly. The item parameter in the Rasch model is called the difficulty parameter and the person parameter is called the ability parameter. The person ability and the item difficulty are expressed in values on the same scale. It is the difference between a person's ability and the difficulty of an item that allows an estimation of the probable outcome when the examinee attempts to answer the item. If the examinee has more ability than is required to solve the item, the probability that he or she will solve the item is greater than the probability that he or she will not. In the case of dichotomously scored items, where the answer can only be right or wrong, the probability of a correct response becomes greater than 50%. How much more than 50% depends on how much 'overpower' the examinee has. The chance of finding the right answer by random guessing is *not* taken into account in the Rasch model. If, however, the examinee has less ability than the item requires, a wrong answer becomes more likely, and the probability of a correct response becomes less than 50%. Figure 1 depicts the relationship between ability and probability of a correct answer to a particular item. This is referred to as the Item Characteristic Curve (ICC).

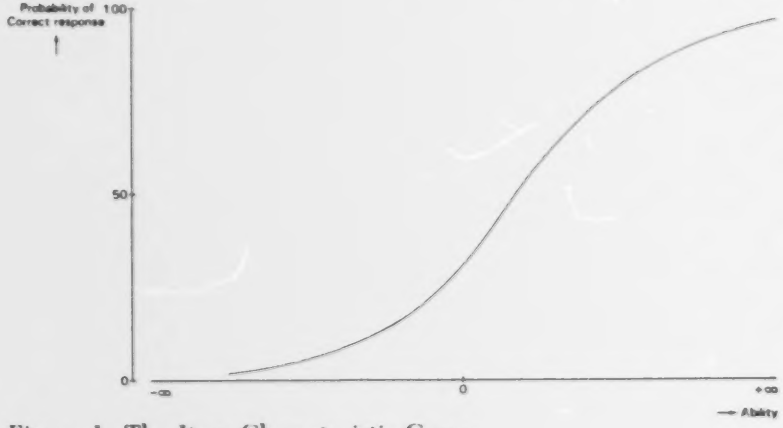


Figure 1. The Item Characteristic Curve

The vertical axis in Figure 1 represents the probability of a correct answer. The horizontal axis represents ability ranging from extremely far below the ability required to solve the item to extremely far above it. At some point to the right, it becomes useless to ask an examinee of that ability to try the item: the item is so easy for the examinee that one 'knows' beforehand that he or she will solve it. In other words, administering this particular item to this particular examinee will not yield new information. On the other side, administering an item that is far too difficult for an examinee will be equally inefficient. To obtain maximum information, the difference between the ability of an examinee and the difficulty of the item should be equal to zero. To maximize the efficiency of the tests at the two MAVO levels, sets of items have to be selected from the pretests that have a difficulty parameter that is equal (or as nearly equal as possible) to the expected ability levels of the two MAVO groups at the date of the final test administration. Figure 2 depicts the Item Characteristic Curves (ICC's) of two items, one for MAVO-C and one for MAVO-D. Below the horizontal axis representing ability, the mean ability and the distribution of ability at the MAVO-C level and at the MAVO-D level is indicated. However, in the pretest situation, MAVO-C and MAVO-D students cannot be distinguished, and the total MAVO group is expected to increase about one logit in ability between the date of the pretest and the final test. Dotted lines indicate the mean ability and the distribution of ability as estimated after the pretest situation for the mixed group of MAVO-C and MAVO-D students. This implies that items selected for the MAVO-C test would have to be of such difficulty that the total mixed-ability group at the time of the pretests would have a probability of .67 of giving a correct answer. For items selected for the MAVO-D test, this probability would have to be .42.

Apart from the appropriate difficulty level, the discrimination power of items at the target level has to be evaluated. In an ideal situation, where all items designed to measure listening comprehension do indeed measure listening comprehension only and in exactly the same way, the construction of tests at different ability levels simply amounts to the selection of a sufficient number of items at the appropriate difficulty level, because all items would discriminate along the same lines and with equal power. However, in reality, measuring listening comprehension is not so simple, and pretests, in particular, cannot be expected to be ideal. On pretests, one can expect to find a number of less-than-perfect items: items whose correct answer does not depend solely on ability in listening comprehension.

Furthermore, the ability to understand a given speech sample may depend on a number of elements of the sample such as range and choice of vocabulary, grammatical complexity, articulation and per-

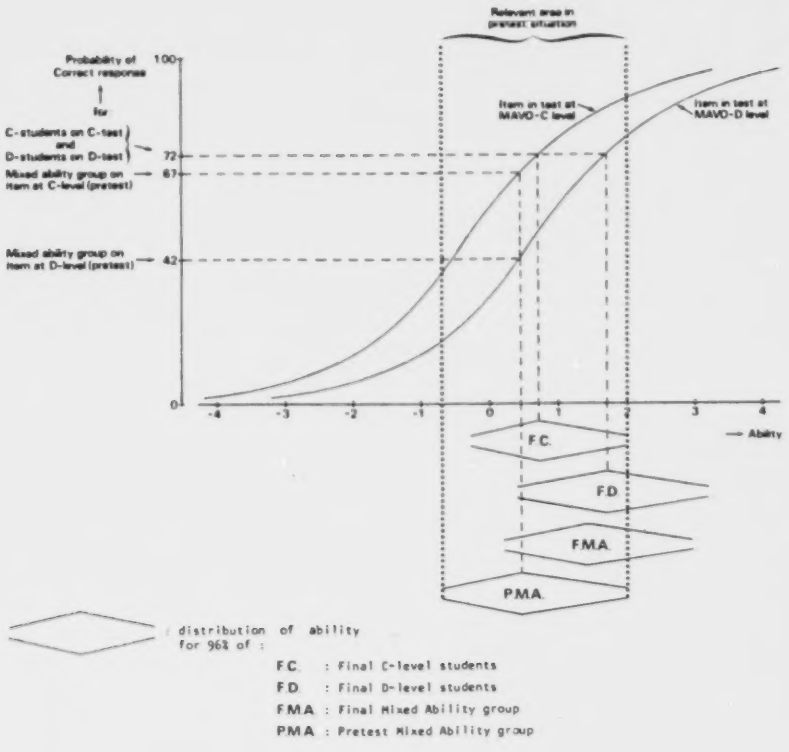


Figure 2. Item Characteristic Curves of two items designed for two different ability groups

sonal speech features of the speaker, abstraction level of the message, density of information, and acoustic conditions. Though a number of these elements can be measured separately, their effect on the difficulty of language – and most certainly the interrelated pattern of influences of these elements – remains uncertain. At the lower levels of ability, these elements are more likely to operate separately than at the higher levels, where the language learner has acquired the competence to make use of acquired knowledge to compensate for lack of knowledge elsewhere. Also, at higher levels of ability the redundancy of natural language enables the listener to infer meaning from a number of elements, and missing one element does not mean missing the whole message. At the lower levels, there may not yet be any possibility for redundancy; the listener is groping for recognizable elements in

the stream of utterances in order to determine what the message is about. Hence items may operate well on one MAVO level because of a certain element but, at the other level, this element may be irrelevant. This irrelevance may be due either to the fact that all students at this level have acquired perfect mastery of the specific element or that all are equally unfamiliar with the element. In either case, the item will lose the power to discriminate with respect to the particular/specific element. However, as this element is only one of a number of elements, another one of the elements may still cause the item to be difficult. It is even possible that items operate discriminatingly at different levels because of different elements. In a given sample of language, students of low ability may or may not be able to understand, for example, the negative in an utterance; those who can will be discriminated from those who cannot, although all may be equally ignorant of the exact meaning of a particular content word. But, at a higher level, all students may recognize the negative; they tend to differ due to knowledge of the content word and it is this knowledge – or the absence thereof – that discriminates between candidates at this higher level. Items like these violate the assumption of unidimensionality and, therefore, do not fit the Rasch model (or any unidimensional model, latent trait or classical). Nonetheless, it may be possible to build a subset of items that does represent a single dimension or trait for one level and a second subset measuring a different dimension for another level.

Therefore, each item pretested on the mixed MAVO group is not only evaluated for its difficulty, but also for its estimated discrimination power at either MAVO level. For this purpose the observed Item Characteristic Curves are inspected for each item. Figure 3 depicts the expected and the observed Item Characteristic Curves (ICC's) for a number of pretest items. In Figure 3, the ability axes are pictured only for the part that is relevant for MAVO students (that is, the part between the dotted lines in Figure 2). Along the horizontal axes, the mean ability of six subgroups from the total MAVO group is indicated. The groups are of equal size. The expected and observed proportion of right answers within each of the subgroups can be read from the vertical axes. Figures 3.1 and 3.2 represent items that, with respect to difficulty, are suitable to be selected for a test at the MAVO-C level. The expected proportion of right answers at the time of the pretest for the total MAVO group is .67 – leading one to expect a proportion of approximately .70 correct answers for the MAVO-C students at the time of the final examinations. Figures 3.3 and 3.4 depict items of the appropriate difficulty for MAVO-D students. The expected proportion of right answers from the total MAVO group on the pretest is .42 leading to an estimated proportion of approximately .70 correct an-

swers from MAVO-D students at the date of the final examinations.

Figure 3.1 depicts the ICC of an item suitable for selection for the final test at the lower MAVO level. It demonstrates that the item discriminates very well and homogeneously at the ability level of MAVO-C students. At the MAVO-D level, however, the item loses discrimination power. The loss is greater than would be expected from ceiling effects. Apparently, there is something in the item that keeps some of the best students from choosing the right answer. At the higher level therefore, the item may cause a homogeneity problem.

Figure 3.2 depicts an item that is also of the right difficulty for MAVO-C students, but it does not discriminate well at this level. For

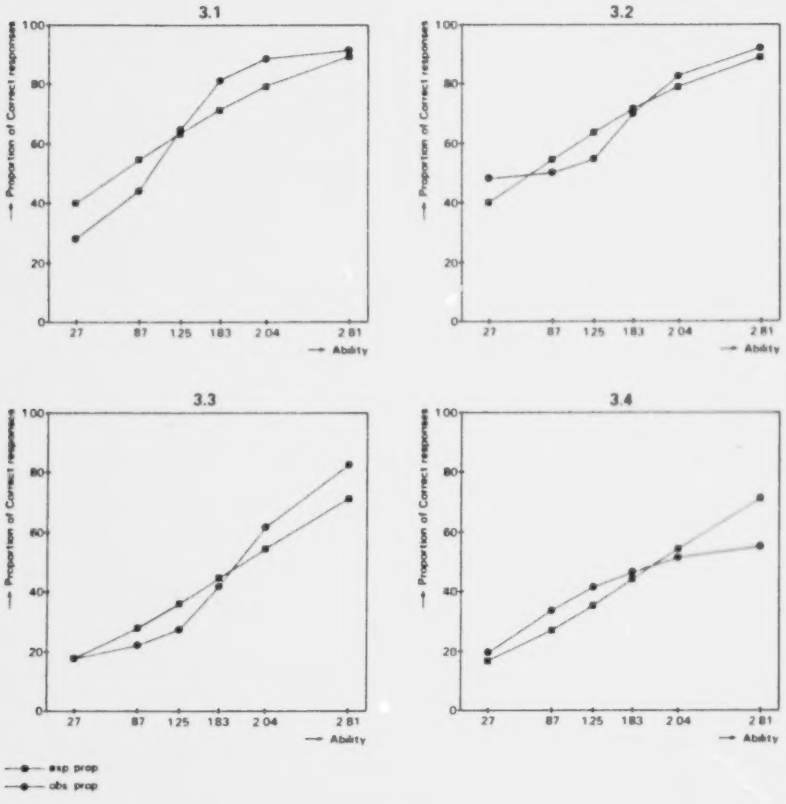


Figure 3. Expected and observed Item Characteristic Curves of four pretest items differing in difficulty and/or optimal discrimination level

students at the MAVO-D level, the item does discriminate, but it is too easy for them. At the time of administration of the final form of the test, the item will have lost discrimination power because of ceiling effects.

Figure 3.3 shows the kind of item that is to be selected for a test at the MAVO-D level. The item is of the appropriate difficulty and discriminates well between students at the higher ability level in the MAVO schools.

Figure 3.4 is an example of an item that, though being at the right difficulty level for the MAVO-D students, is ineffective because it does not discriminate between students at this level.

Evaluation

To evaluate the effectiveness of the described item selection method, inspection of data from tests of the first 3 years can be used. The introduction of a single curriculum with two different examination levels for MAVO schools was a political decision. Solving the problem of testing mixed-ability groups was left to the Dutch National Institute for Educational Measurement (CITO). There was no time to create and evaluate methods of test development for this new situation before the new system was introduced. The method described in this paper was developed on the basis of retrospective studies of examination results during the two years following its introduction. The effectiveness of a method for selecting items from pretests can be evaluated using data from the administration of the final version of the tests.

As discussed earlier in this paper, students decide at which level they will sit for the external part of their examinations on the basis of their results on the internal parts. The listening comprehension test is one of the internal parts. At the time of the internal examination, students do not yet have to decide which level of the external examination they will take. Therefore, all MAVO students sit for the listening comprehension tests at both levels. Several months after the listening comprehension tests have been administered, students decide at which level they wish to graduate and, consequently, which of the two listening comprehension tests will be used to determine their mark. Data on the final level of the students have been collected for the 1983 and 1984 tests. These data can be used to evaluate the actual appropriateness of the tests for the levels for which they have been designed. The method described in this study was developed on the basis of this evaluation and was applied in selecting items from pretests of the 1985 listening comprehension tests. While these tests were administered in February 1985, final decisions of the students with respect to their level will not be known until September. However, at this time, we can estimate the effectiveness of the method based on:

(a) a comparison of "predictions" as to the appropriateness of items with the actual results of the two subgroups on the final version of the 1984 tests;

(b) a comparison of the 1983 and 1984 results with the 1985 test results as they appear from the total mixed ability groups.

In fact, the "predictions" on the 1984 tests were made after the final choice of the students was known, but without making use of this information. The items in the 1984 test had been selected from the pretests mainly on the basis of their estimated difficulty. A total of 75 unaltered items were selected to be included in the final version of the MAVO tests. Rasch analysis of the pretests showed that for half of these items, the ICC was steeper at the level of MAVO-C students; for the other half, the ICC was steeper at the level of MAVO-D students. Table 1 presents the relative height of the r_{ir} of these items in the definitive version of the tests at the two MAVO levels related to the ICC of the same items in the pretest on the total mixed-ability group. (The r_{ir} of an item is the correlation of this item with the rest of the items in the test. It is considered a discrimination index.)

Table 1
Relation between Item Characteristic Curve in Pretest and r_{ir} in Definitive Test in Two Different Ability Groups

ICC in pretest steeper at	highest r_{ir} in definitive test on lower group	higher group	totals
lower ability range	23	15	38
higher ability range	5	32	37
totals	28	47	75

Index of predictive association (Goodman & Kruskal, 1954): .29 Chi square: 21.56, df. 1, $p < .005$

In general, items have higher correlation indices when the subjects taking the items are distributed more widely in ability. The MAVO-D students form a much larger and more differentiated group than the MAVO-C students. Therefore, items are more likely to correlate highly when the tests are administered to the MAVO-D students. Even without taking this expected higher correlation into account, the relation between the ICC in the pretest and the r_{ir} in the final test is significant.

Table 2 summarizes the results of a sample of 1,000 students from the total (mixed ability) group of MAVO students on the 1983, 1984, and 1985 listening comprehension tests for English, French, and German. The test length (k), the mean proportion of correct answers (p), and the reliability (K R 20) are given. Because the tests vary widely in length, the reliability index is reestimated using the Spearman-Brown formula at a uniform length of 50 items. This is included in the table.

Table 2
Summary of Test Results Total Mixed Ability Groups

	1983				1984				1985			
test	k	p	KR20	(*)	k	p	KR20	(*)	k	p	KR20	(*)
MAVO-C												
English	40	.83	.81	(.84)	37	.79	.81	(.85)	43	.80	.80	(.82)
French	50	.84	.70	(.70)	44	.84	.65	(.68)	38	.80	.71	(.76)
German	40	.79	.82	(.85)	38	.84	.79	(.83)	44	.80	.80	(.82)
mean	43	.82	.78	(.80)	40	.82	.75	(.79)	42	.80	.77	(.80)
MAVO-D												
English	40	.79	.81	(.84)	50	.70	.80	(.80)	47	.70	.83	(.84)
French	38	.72	.69	(.75)	40	.76	.73	(.77)	42	.67	.79	(.82)
German	40	.73	.75	(.79)	42	.75	.81	(.84)	47	.69	.83	(.84)
mean	39	.75	.75	(.79)	44	.74	.78	(.80)	45	.69	.82	(.83)

k: test length; p: mean proportion of correct answers;
KR20: Kuder and Richardson reliability index;
(*): Spearman-Brown prediction at test length of 50 items

In 1985 the tests for the three languages differ less in mean p-value than in former years when the item selection method was not used. This indicates that in 1985 the difficulty of the tests for all three languages is equally well-targeted on the ability of the students. The reliability (K R 20) tends to be more uniformly high in 1985 – as can be

expected from tests that are better on target. Since the largest proportion of the total mixed-ability group eventually chooses to graduate at the higher level, the tests for the higher level (MAVO-D) are expected to measure more accurately than the lower-level tests (MAVO-C) when administered to the total mixed-ability group. This is particularly true for the 1985 tests.

In Table 3, estimated mean proportions of right answers on the MAVO-C and MAVO-D tests from the two separate target-groups (MAVO-C and MAVO-D students) are presented for the 1983, 1984, and 1985 tests. The estimates are based on the observed mean p-value in the total mixed-ability group (cf. Table 2) and on the expected distribution of MAVO students over the two subgroups. In addition, the observed values of the 1983, 1984, and 1985 tests are given. The observed values for the 1985 tests only recently became available.

Table 3
Estimated and Observed Proportion of Correct Answers at the Intended Ability Levels

	1983		1984		1985	
test	Estim.	Observ.	Estim.	Observ.	Estim.	Observ.
MAVO-C						
English	.70	.67	.71	.68	.73	.69
French	.77	.77	.77	.77	.73	.70
German	.69	.63	.75	.70	.73	.72
MAVO-D						
English	.77	.76	.73	.73	.73	.73
French	.75	.75	.79	.79	.71	.72
German	.76	.76	.78	.77	.72	.72
mean	.74	.72	.76	.74	.73	.71
stand. dev.	.03	.06	.03	.04	.01	.01

The estimates correspond quite well with the observations, as can be seen from the 1983, 1984, and 1985 data. As the proportion of students opting to graduate at the lower level is relatively small (15 to 18%), the estimates for the lower-level tests tend to be less accurate than the estimates for the higher-level tests. In 1985 the tests appear to be better targeted on the subgroups, as can be seen from the fact that the estimated mean proportions of right answers in both subgroups are closer together in 1985.

Finally, if the tests do measure at different levels of ability and are to help students decide at which level they will sit for the external examinations, the lower-level test should be easier than the higher-level test for every student; each student should have a higher probability of giving the right answers on the lower-level test. But, as the lower-level test is meant to measure more accurately at the lower level, measurement error is expected to increase with ability and, subsequently, the chance of obtaining a higher or equal score on the higher-level test will go up with increasing ability. A student obtaining a perfect score on the higher-level test is also expected to obtain a perfect score on the lower-level test (because no higher score is available), and so this student will have the same score on both tests. Figure 4 depicts the relationship between the proportional scores for samples of 1,000 students from the total MAVO (mixed-ability) population on the higher-level test (MAVO-D) and on the lower-level test (MAVO-C) of listening comprehension of German in 1983 and in 1985. Students obtaining better or equal results on the higher-level test than on the lower-level test are plotted as a proportion of all students obtaining a particular score on the higher-level test. The cutoff scores for 1983 and 1985 are indicated by an arrow. In the 1983 and 1985 tests, the probability of obtaining a higher or equal score on the higher-level tests increases with ability, as expected. In 1983, however, quite a large proportion (± .20) of the students who obtained scores at the critical level (from very low scores to just above the cutoff score) did better on the higher-level test, which indicates that these students of relatively moderate ability had a better chance to pass by taking the test at the higher level. In 1985 only a marginal proportion of students in the critical area obtained unexpected higher scores on the higher-level test.

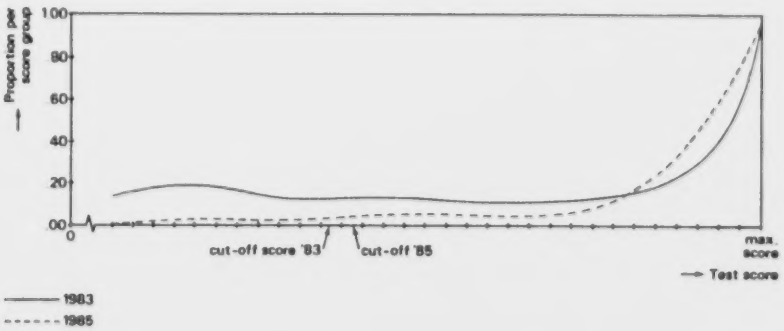


Figure 4. Students scoring higher or equal on higher level test: German listening comprehension tests 1983 and 1985

Though it is too early to pronounce a definitive judgment on the proposed item selection method, it can be concluded that the results so far are promising. Research on adaptive or tailored testing indicates that tests can be designed to yield maximum information about students at predetermined ability-levels. The method discussed here helps to get maximum information about items from students at undetermined ability-levels.

References

Boeijen, M. (1984). Testing listening comprehension: notions and functions as "discrete points" in a communicative context. In T. Culhane, C. Klein Braley, & D. Stevenson (Eds.), *Practice and problems in language testing*, Occasional Papers No. 29. Colchester: University of Essex, Department of Language and Linguistics.

De Jong, J. H. A. L. (1983). Focusing in on a latent trait: an attempt at construct validation by means of the Rasch model. In J. van Weeren (Ed.), *Practice and problems in language testing 5: non-classical test theory. Final examinations in secondary schools*. Arnhem, Netherlands: Centraal Instituut voor Toetsontwikkeling (CITO).

De Jong, J. H. A. L. (1984). Testing foreign language listening comprehension. *Language testing, 1*(1), 97-100.

De Jong, J. H. A. L. (1985). Effect of mixing ability groups on ability levels attained. In V. Kohonen & A. J. Pitkanen (Eds.), *Language testing in schools*(AFinLA Series 41). Tampere, Finland: Association Finlandaise de Linguistique Appliquee (AFinLA).

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications*. New York: Springer-Verlag.

Groot, P. J. M. (1975). Testing communicative competence in listening comprehension. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency*. Washington, DC: Center for Applied Linguistics.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin, (Eds.), *Test equating*. New York: Academic Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Van Ek, J. A. (1975). *The threshold level*. Strasbourg: The Council of Europe.

Wright, B. D. & Mead, R. J. (1975). *CALFIT*. (Research Memorandum No. 18). Chicago: University of Chicago, Department of Education.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: Mesa Press.

SECTION TWO

Using technology to develop new measures of speaking, reading, and writing.

Correlation between Human and Machine Ratings of Test of Spoken English Reading Passages

Garry Molholt and Ari M. Presler
Rensselaer Polytechnic Institute, Troy, New York

Abstract

This paper describes a pilot study of the feasibility of computer-assisted evaluation of pronunciation. Twenty samples of speech were ranked by professionals in the Test of Spoken English (TSE) program at Educational Testing Service (ETS). Subjects read the same passage from a disclosed form of the TSE. The passages were analyzed on a Harris 800 super minicomputer at the Speech Processing Laboratory of Rensselaer Polytechnic Institute. Because the machine-generated scores ranked the passages in nearly the same order as the human raters, this study indicates that machine scoring is feasible.

Correlation between Human and Machine Ratings of Test of Spoken English Reading Passages

Purpose of Machine Rating of TSE Reading Passages

There are two major reasons why machine rating of oral proficiency is attractive. They are:

1. Machine rating is much more efficient than human rating, especially when dealing with large amounts of data.
2. Machine rating could be used to identify and classify weak areas of pronunciation as part of a system of computer-assisted instruction in pronunciation.

This paper focuses on the first of these advantages.

According to Pendergast (1980, p. 101), it would take two full-time interviewers working all day every day to provide reasonable feedback on oral proficiency for a program for 250 students. For large scale operations, this ratio would be extremely expensive. By dividing the tasks between those that must be done by people and those that may be done as well or better by machines, we propose to greatly reduce the number of people needed, while increasing the quality of the evaluation.

Basic Characteristics of Speech Measurement

Do you know how a microphone works? Inside there is a diaphragm with an attached sensor. As sound waves vibrate the diaphragm, the sensor sends a signal. This signal has three properties: frequency, amplitude, and duration. Frequency refers to the number of times the sensor moves in a second. Amplitude refers to how far it moves. These terms are exemplified in Figures 1 and 2. Sounds emanating from human speech may be characterized by frequency, amplitude, and duration. They are not, however, simple waves as in Figures 1 and 2. They are complex, as in Figure 3. The waves are complex because several components join together to form a speech sound. To fully analyze speech sounds, they must be broken down into their component parts and compared. The wave in Figure 3 is composed of

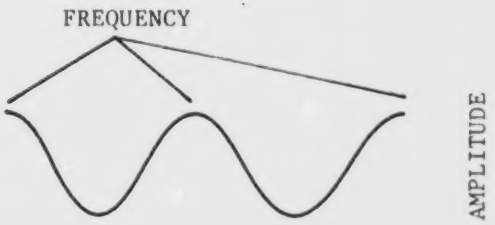


Figure 1. Low frequency, high amplitude

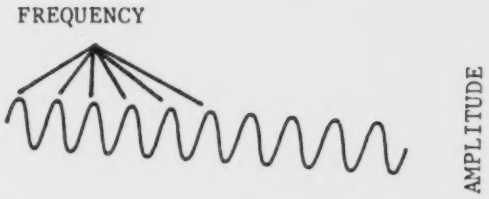


Figure 2. High frequency, low amplitude

Figure 3. Complex wave

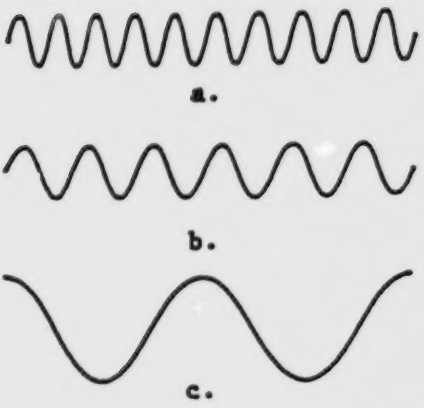


Figure 4. Components of a complex wave

Components a, b, and c of Figure 4. Component c of Figure 4, with the lowest frequency, is called the fundamental frequency. All components above the fundamental frequency are harmonics of that frequency. Many speech sounds may be differentiated by recognizing which harmonics of the fundamental frequency have high amplitude.

Brief History of Electronic Analysis of Speech

The vowels of American English can be easily differentiated according to the relationship between their fundamental frequency and the next higher level. This is shown in Figure 5. The horizontal axis shows the fundamental frequency (first formant), and the vertical axis shows the next higher level (second formant).

Thus /i/ has its first formant between 200 and 400 Hz. and its second between 2000 and 3500, whereas /u/ has a similar fundamental frequency, but its second formant is only between 600 and 1200 Hz. These relationships were shown clearly by Potter, Kopp, and Kopp (1966, p. 22). Figure 6 is a vowel chart from their text. It shows bands of frequency that have high amplitude. Note that the second band for /i/ is much higher than the second and for /u/. This corresponds to the information in Figure 5.

Early speech scientists recognized that consonants are not as easy to analyze as vowels. Voiceless stops, such as /p/ and /t/ were displayed as blank spaces, as in Figures 7 and 8, which are taken from Rabiner and Schafer (1978, p. 60).

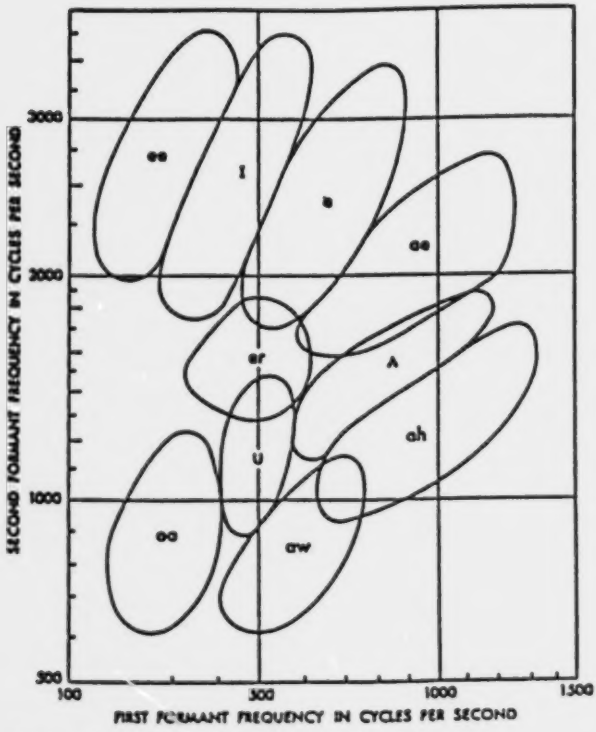


Figure 5. Vowels by first two formants

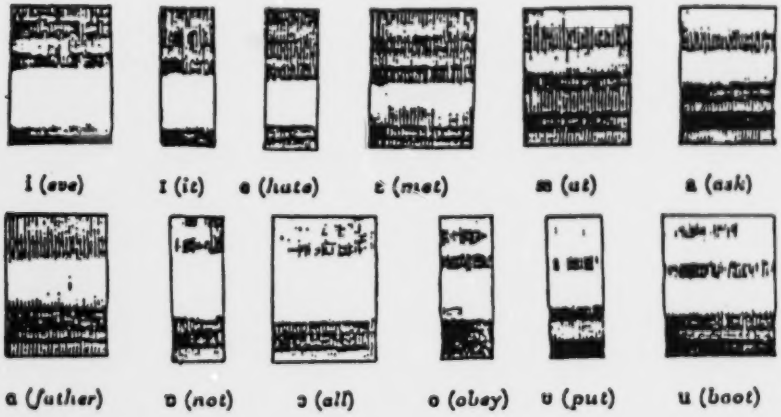


Figure 6. Vowel frequencies

BEST COPY AVAILABLE

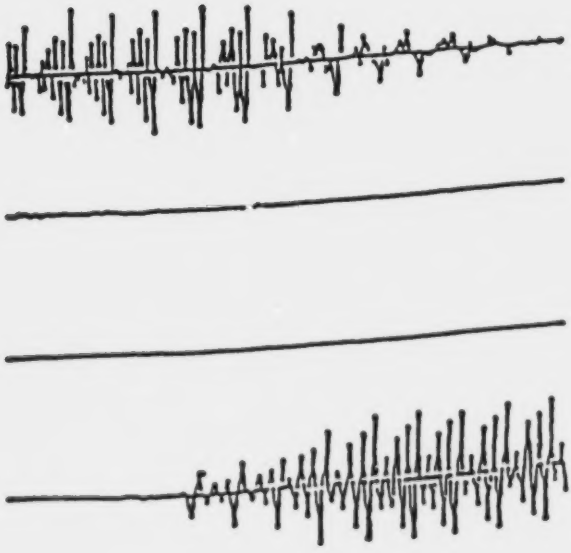


Figure 7. epa in time domain

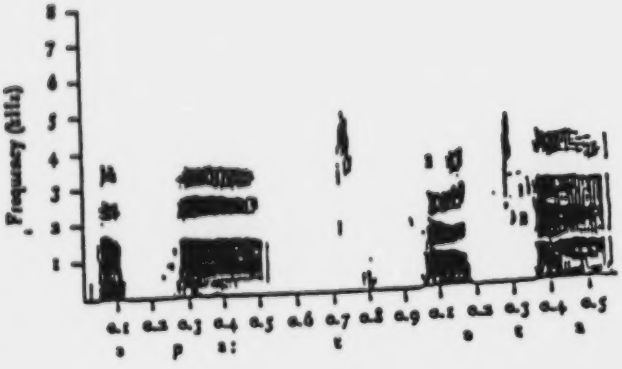


Figure 8. epa and eta frequencies

It was later recognized that different consonants have different effects on vowels. For example, /b/ causes the second formant of a following /i/ to start much lower than normal and then slide up to its level. With /g/, the second formant of /i/ starts much higher than usual and slides down to its normal level. These are shown in Figure 9. By looking at the way the vowels are disturbed, it is possible to differentiate the consonants.

Though the discovery of this technique for analyzing speech represented a major breakthrough, it was not economically feasible to put it into wide use until recently. In the past computers were not as powerful or as affordable as now. Today there are many sophisticated techniques of analysis. As electrical engineers strive to perfect voice activated machinery, linguists have a new opportunity to learn which techniques might be useful if applied to language analysis and education. A potential area for the application of speech processing techniques to teaching English as a second language (TESL) is the evaluation of pronunciation.

Research Conducted for this Study

This paper describes a small feasibility study using a Harris 800 super minicomputer to score and rank the pronunciation of 20 nonnative speakers of English reading the same passage from a disclosed

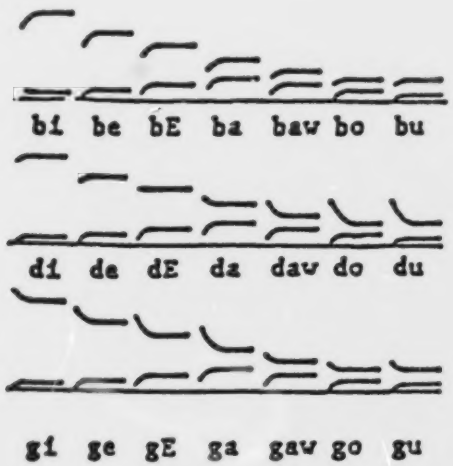



Figure 9. CV Transitions

form of the Test of Spoken English (TSE) published by Educational Testing Service (ETS).

A tape containing speech protocols of 20 examinees who took the TSE was provided by ETS staff. Each protocol contained the response to the reading aloud portion of the test. The 20 subjects were chosen to represent a broad spectrum of linguistic backgrounds and abilities in English. Each recorded the following passage on tape as part of his or her examination performance.

> An increase in the number of visitors to many of the natural wonders in the United States is a growing problem. One example is the Grand Canyon in Arizona. Twenty years ago, the number of visitors averaged 2 million each year. Today, the average number is 3 million and, during the summer months, as many as 30 thousand people a day crowd onto the southern rim of the Canyon. Even the so-called wilderness areas, such as the Canyon floor, are becoming congested. In order to solve the traffic problems created by this concentration of people, cars have been banned from the major highways and the people are brought to the park in buses. Officials predict that in another three to four years all entrances will have to be closed as early as noon.

Through the use of the computer, this text can be represented by a set of shapes composed of a sequence of energy waves. For example, the digit *three* exhibits the shape of a voiceless fricative /th/ followed by a semivowel /r/ and the vowel /i/, as in Figure 10. If a speaker fails to

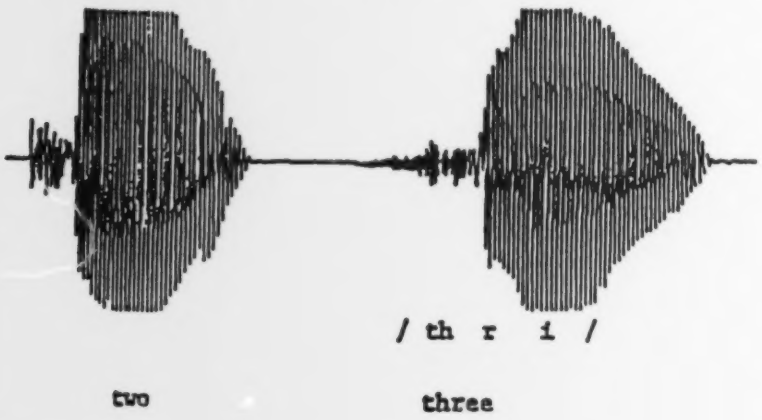


Figure 10. Three, spoken correctly

produce the fricative, the result may look like Figure 11. The difference between the shapes of /th/ in Figure 10 and unaspirated /t/ in Figure 11 is easy to display with a computer. Therefore, it is also easy to analyze.

The features extracted by the analysis system precisely represent changes in the shape of up to 10 areas of the vocal tract plus voicing and turbulence. To evaluate a speaker's pronunciation, his or her speech is matched against a reference template. A reference template contains the average of a broad or narrow group of speakers pronouncing the same sounds as the tested speaker. Any deviation in the test speech from the speech in the reference template is noted as having an improper distance from the expected curve. The cutoff distance used depends on the application. For speaker identification the cutoff is very low. For comprehensibility it is much higher.

Characteristics of tested speech can be evaluated at the level of phonemes, syllables, words, phrases, sentences, or even voice types. For this research, the phoneme level was used with a relatively high cutoff distance. Measuring comprehensibility is the main goal. Since a program of automatic segmentation of phonemes is still in the developmental stages, phonemes were segmented by entering coordinates from voice prints of the speech files. This provides a simulation of a completed system.

Errors which provide significant deviation from the expected curves include:

0 = missing phoneme
X = wrong phoneme
\- = partial phoneme
\+ = extra phoneme
! = improper stress

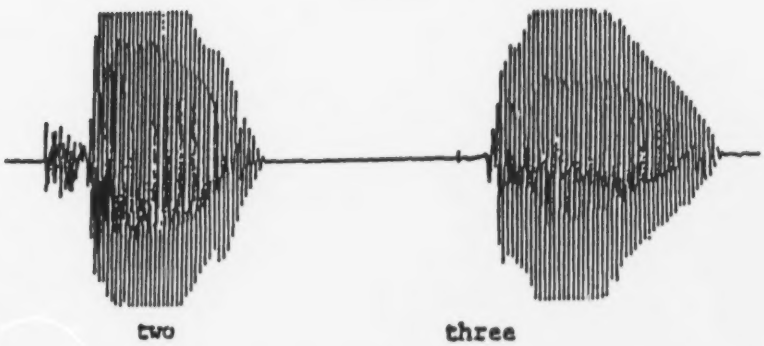


Figure 11. Three, spoken incorrectly

The analysis programs recognize and tabulate the occurrence of each error. This provides the means for ranking speakers in terms of comprehensibility. Table 1 provides a list of the overall comprehension scores for the passages as rated by TSE professionals and the number of significant errors for each passage as determined by this research. The rank order correlation coefficient (Spearman's Rho) is .9318. This compares quite favorably to the .88 intercorrelation between pronunciation and fluency reported in the *Test of Spoken English Manual for Score Users* (1982, p. 23).

Table 1

SAMPLE	OVERALL TSE SCORE	MACHINE SCORE
1	110	-67
2	150	-55
3	160	-29
4	180	-26
5	190	-18
6	200	-18
7	200	-12
8	200	-14
9	210	-12
10	220	-9
11	220	-11
12	240	-8
13	240	-7
14	250	-6
15	250	-5
16	260	-5
17	230	-3
18	230	-5
19	280	-5
20	280	-1

Examples from the TSE Passages

The following figures show comparisons of tested subjects with weak pronunciation and strong pronunciation. For the convenience of this description, displays are in the time domain, with amplitude on the vertical axis and duration on the horizontal axis. It is easier to see the diversions of syllables and words in this type of display than in the frequency domain, with frequency on the vertical axis.

Figures 12 and 13 show an incorrect and a correct pronunciation of *created*. Note that /k/ is a high frequency low amplitude sound; thus it has many little curves crossing the center line. The vowel sound has much higher amplitude. The final voiced stop has gradually decreasing amplitude over a relatively low frequency. Shapes occurring in proper sequence represent speech that can be easily understood. Figure 13 shows a correct version of *created*. It has all the proper shapes in the proper sequence.

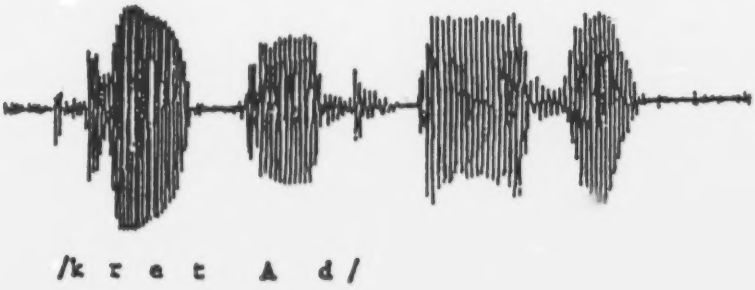
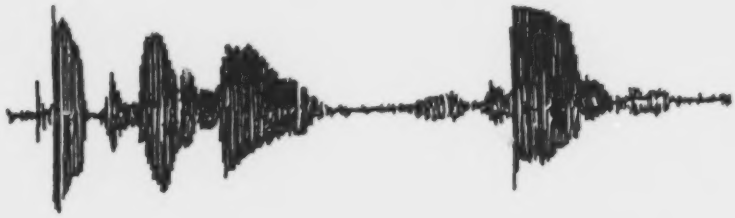


Figure 12. created (missing /i/)

Figure 13. created (correct)

Within only 9/10 of a second, the speaker represented in Figure 14 lost 5 points. *Cars*, which should be /karz/ is pronounced as /kae/. The shapes for both the semivowel and voiced fricative are missing. The vowel is also wrong. This would be easier to see in the frequency domain. *Been* is pronounced with too much stress. It should not have greater amplitude than *banned*, which is missing its word final voiced stop.

In Figure 15, all the correct shapes are present in the proper sequence for *cars have been*. While Figure 16 depicts a correct representation of *banned*, Figure 17 shows an incorrect version, with the silent vowel of the past tense marker pronounced as if the verb stem ended with /t/ or /d/. Thus, the word in Figure 17 exhibits an extra phoneme.



/k ae v bEn b ae n /

Figure 14. cars have been banned (incorrect)

/ k a r z h ae v b E n /

Figure 15. cars have been (correct)

Figure 16. banned (correct)

Figure 17. banned (incorrect)

The next two Figures represent the phrase *traffic problems*. Figure 18 shows that the second syllable of *traffic* is much too weak and is missing the final voiceless stop. The final /z/ of *problems* is also missing. In Figure 19 all shapes are proper.

As can be easily seen from these examples, each class of sounds, such as voiced stops or voiceless fricatives, has a unique and characteristic shape in the time domain. Thus a system that recognizes shapes in sequences is invaluable for the evaluation of pronunciation. Most errors seem to occur between classes of sounds, not within classes of sounds. This may explain why pronunciation books stress the differences between /b/ and /p/ but not /b/ and /d/. They drill /v/ and /f/, but not /v/ and /z/.

Figure 18. traffic problems (incorrect)

Figure 19. traffic problems (correct)

With additional information from the frequency domain, especially for vowels, we have a system that accurately characterizes the level of pronunciation.

Equipment Specifications

Normally, the process to achieve the above results is long and arduous. However, with the use of the Harris H800 super minicomputer, the process is reduced to a few simple routines. Functioning at a rate of almost 4 million instructions per second (IPS), the computer instructs the peripheral to measure the energy level of the speech 10,000 times a second. This analog to digital device (A/D) was developed by Gruber (1984). It inputs information from various sources such as computer tape, audio tape or microphone. In turn, it generates

a file that contains the information to be further processed through the software. The current configuration allows processing of up to 10 seconds of speech for each file, providing 100,000 pieces of information. This information is then analyzed through mathematical models including linear predictive coding (LPC) to extract the essential characteristics of speech (Greub, 1985, p. 4).

The uses and applications of such equipment are not limited to those presented, but can easily be applied to other areas, including communication and security. Several areas of application now being studied further include recognition of voice patterns, speaker verification, and data compression.

References

Educational Testing Service. (1982). *Test of spoken english: manual for score users.* Princeton, NJ: Author.

Greub, H. (1985). *Speech recognition system.* Unpublished master's thesis, Rensselaer Polytechnic Institute, Department of Electrical, Computer, and Systems Engineering, Troy, NY.

Gruber, C. (1984). *Implementation of a linear predictive coding analysis/synthesis system for speech compression on a newly established computer facility.* Unpublished master's thesis, Rensselaer Polytechnic Institute, Department of Electrical, Computer, and Systems Engineering, Troy, NY.

Pendergast, T., Jr. (1980). OLAF N. 73: A Computerized Oral Language Analyser and Feedback System. In Y. P. Lee, A. Fok, R. Lord, & G. Low (Eds.), *New Directions in Language Testing.* New York: Pergamon Press.

Potter, R. K., Kopp, G. A., & Kopp, H. G. (1966). *Visible Speech.* New York: Dover Publications.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice Hall.

Authors' Notes

The authors would like to thank Professor Lester Gerhardt, Chairman of the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute for his kindness, encouragement, and support. We also thank Miss Hsing-Huei Huang for assisting in the development of the analysis process.

Development of a Tape-Mediated, ACTFL/ILR Scale-Based Test of Chinese Speaking Proficiency

John L. D. Clark
Center for Applied Linguistics, Washington, D.C.

Abstract

This paper describes the design, development, and initial administration of a tape-mediated "semi-direct" test of Chinese speaking proficiency for American and other English-speaking learners of Chinese. This test – which is designed to be scored and interpreted according to the ACTFL/ETS/ILR guidelines for direct proficiency tests, but also to be usable in situations where actual face-to-face testing is financially or administratively infeasible – makes use of several different elicitation techniques, including two-way "conversations"; picture-based narration; description of personal background and interests; and performance of a variety of situational tasks, including tasks requiring differing linguistic registers and affectively marked discourse.

Considerations in the selection of these particular question types by comparison to other possible formats are discussed, as well as the measurement advantages and disadvantages of tape-based administration by comparison to the live interview. Results of the administration of a preliminary form of the test are presented, together with a discussion of technical and practical aspects of test development for tape-based instruments, and recommendations for further research and validation activities in this area.

Development of a Tape-Mediated, ACTFL/ILR Scale-Based Test of Chinese Speaking Proficiency

Foreign language instruction in the United States may be considered to be in the early stages of a major revolution, a revolution that addresses, simultaneously and in a highly synergistic manner: the intended outcomes of language instruction, the pedagogical materials and procedures used to bring about these outcomes, and the assessment techniques most validly employed to certify their attainment. The present status of and anticipated further developments within the "proficiency" movement, as this broad initiative is coming to be called, are well summarized in several recent publications, including especially three recent volumes of the ACTFL Foreign Language Education Series (Higgs (Ed.), 1982; Higgs (Ed.), 1984; James (Ed.), 1985) and the winter 1983 special issue of the *Modern Language Journal* (Ervin (Ed.). 1983) entitled "Pathways to Change in Language Teaching."

With regard to the measurement component of the proficiency movement, a major recent undertaking – through the joint efforts of the American Council on the Teaching of Foreign Languages (ACTFL), Educational Testing Service (ETS), the CIA Language School, the Defense Language Institute (DLI), and several other member agencies of the U. S. government's Interagency Language Roundtable (ILR) – has involved a thoroughgoing review of the original Foreign Service Institute (FSI) Absolute Proficiency Ratings.[1] This review has resulted in relatively minor changes in the original broad level descriptions, but has, very usefully, brought about the incorporation of appreciably more detail in the full "paragraph" expansions of each of the level descriptions. The revised level descriptions, together with the detailed verbal expansions, have been adopted, as of December 1983, as the official, government-wide descriptions of language proficiency for individual student assessment, training program evaluation, proficiency-based salary increments within given employment categories, and other purposes. At approximately the same time as the ILR definitions were undergoing review, collaborative efforts were being made by ACTFL and ETS (see Liskin-Gasparro, 1984) to prepare and incorporate into the original proficiency scale a number of finer-grained descriptive levels within the 0-3 range of the total scale.

Provision of these additional scale descriptions was intended to permit more sensitive measurement of student performance within that portion of the total score range usually at issue in academic-based language instruction at the secondary or college level.

The development and refinement of the ILR proficiency guidelines; development and dissemination of the ACTFL-ETS "expanded" descriptions; and the subsequent carrying out, by ACTFL, ETS, and other organizations, of numerous familiarization sessions and intensive training workshops on interview-based testing and related curricular implications augur well for the continued provision of the numerous benefits associated with the "proficiency" movement for those languages for which the total student volume, number of qualified instructors, and other material and personnel resources are adequate to support this type of curriculum development and assessment. However, for a number of the so-called "less-commonly-taught" (LCT) languages – usually considered to include all world languages other than English, French, German, Italian, Portuguese, Russian, and Spanish – the relatively small number of students involved with the language on a nationwide basis (and the associated very modest course enrollments at individual institutions) make it difficult or impossible to support financially the testing component of the overall process, that is, to defray the costs of local training of interview testers/raters or, alternatively, to support the travel costs and other expenses involved in bringing in outside evaluators for student testing purposes.

It is the intent of a project under way at the Center for Applied Linguistics (CAL), for which progress to date is reported here, to develop an alternative technological approach for speaking proficiency testing that will be operationally practical for use with the lower-volume languages and, at the same time, highly congruent with the regular interview-based ACTFL/ILR testing procedure with respect both to the types and degrees of language performance measured in the test and the reporting and interpretation of test results. The major sub-objectives of this project are: (a) to develop four alternative forms of a "semi-direct," tape- and booklet-mediated test of Chinese speaking proficiency that can be administered in a highly automated fashion either in a language laboratory or in a regular classroom using portable tape recorders; (b) to investigate the validity of the semi-direct approach as a useful and reliable surrogate for the "live" interview for situations in which direct face-to-face testing is not feasible; and (c) to prepare a detailed report of project activities in a form that will provide useful information and procedural recommendations for other individuals or groups considering the development of similar instruments for other LCT languages. As of the present time, one preliminary form of the Chinese Speaking Test (CST) has been de-

veloped and administered, on a clinical basis, as a model for the four final versions of the test, currently in preparation.

Conceptual and practical advantages and disadvantages of "semi-direct" testing, defined as any procedure which elicits active speech by the examinee in response to tape recorded stimuli, printed test booklets, or other "non-human" means, in comparison to "direct" face-to-face testing have been discussed in detail in Clark (1979). For certain applications, such as the diagnostic testing of the examinee's control or lack of control of discrete lexical or structural points, the semi-direct approach may be considered the procedure of choice, in that it permits the systematic presentation of highly focused and unambiguous stimuli in a way that is not generally feasible in a live conversational setting. By contrast, a major disadvantage of semi-direct techniques for general proficiency testing is their lack of face validity as work-sample measures of the examinee's ability to use appropriately a variety of discourse management strategies of major importance to effective face-to-face communication, including, for example, following appropriate turn-taking conventions, requesting clarification as necessary, repairing instances of miscommunication, and so forth. It is not the intent of the current project to attempt to produce a semi-direct test that will validly assess these and other discourse-related aspects of the examinee's performance (such as undertaking would be destined to failure from the start), but instead to measure as validly and as efficiently as possible those numerous other aspects of oral performance that do not hinge on discourse management considerations per se but constitute an important and integral part of overall speaking proficiency as this concept is understood and used in the ACTFL/ILR context.

Previous Semi-Direct Testing Activities

A fairly large number of "semi-direct" tests of foreign/second language speaking ability have already been developed and used for a variety of purposes. Such instruments include, for example, the early MLA-Cooperative Speaking Tests, the tape- and booklet-mediated speaking tests produced under the auspices of the ETS Advanced Placement Program, and the Test of Spoken English, recently developed by ETS as an oral complement to the listening comprehension and reading sections of the Test of English as a Foreign Language (TOEFL) (Clark & Swinton, 1979). For purposes of further discussion, it will be helpful to use a somewhat expanded term "proficiency-oriented semi-direct test" to refer to any semi-direct test that is intended to (a) approximate as closely as possible the linguistic content and manner of operation of a live interview; and (b) be scored and interpreted on the basis of the ACTFL/ILR, or functionally similar,

general proficiency guidelines.

To the writer's knowledge, prior to the current project, only a single test development effort had taken place that could be considered to fall into the "proficiency-oriented semi-direct" category as defined above: the Recorded Oral Proficiency Examination (ROPE), developed and reported by Lowe and Clifford (1980). Versions of the ROPE have been prepared for Chinese and several other languages, but availability is restricted to administration by government agencies on a secure basis. The test consists of a series of tape-recorded questions in the target language to which the examinee replies orally, with responses tape recorded for later evaluation. In a validation study conducted by the test developers, quite high product-moment correlations ($r = .90$) were obtained between examinee ratings on the ROPE and on a concurrently administered face-to-face interview. These results, while certainly not conclusive or procedural or interpretive isomorphism between the two types of instruments, did provide some degree of encouragement to project staff about the prospects of developing a semi-direct instrument whose overall statistical performance would approximate fairly closely that of the live interview.

Considerations in Test Format Selection

Format specifications for the Chinese Proficiency Test were developed in the fall of 1984 by a working committee of CAL staff, Chinese language experts, and other resource persons highly familiar with the ACTFL/ILR interviewing procedure and associated rating system. In the course of this initial planning phase, the committee discussed in detail and came to general agreement on six major content/procedural requirements which the test would be expected to meet, as indicated below.

1. *The test should be designed to cover a range of proficiency from 1 to 3+ (Intermediate Low to Superior)[3] on the ACTFL/ILR scale.*

Very early in the test format discussions, it was acknowledged that it would not be possible or desirable, for both practical and measurement-related reasons, to attempt to develop an instrument that would cover the full 0-5 range of the oral proficiency scale. At the upper end of the scale, in view of the exponential increase in range and sophistication of speech associated with each higher rating level, it was considered impossible to elicit, in a test of any reasonable length (and without the live-interview capacity to interactively tailor the stimulus probing), a sufficiently extensive speech sample to reliably determine level 4 or higher proficiency, or to adequately differentiate examinee

performances within the 4, 4+, 5 range. From a practical standpoint, holding the "top end" of the test to the 3+ level was not felt to hinder to any appreciable extent the effective use of the test within typical U.S. academic programs, for which the proficiency outcomes are generally not beyond the 3+ level.

At the lower end of the scale, a level 1 cut-off was established for two reasons. First, it was felt that in order to validly assess performance below the "survival" level – the level at which the examinee first demonstrates minimal functional proficiency in real-life communicative settings – it would be necessary to use rather artificial elicitation procedures (for example, asking the examinee to identify colors, name the objects in a room). These procedures would have little or no face validity as reflective of real-life communication tasks and would be both conceptually and operationally out of keeping with the other, more highly face valid portions of the test. A second reason for not including "sub-level 1" material was to avoid the possibility of suggesting to potential test users that the ability to enumerate lexical items within specified categories, or in other ways manipulate language in a less than functionally communicative manner, constituted an acceptable demonstration of "proficiency" for real-life language-use purposes.

2. *To provide the greatest possible flexibility in stimulus presentation, the test should make use of a printed test booklet in addition to a master audiotape.*

In the ROPE, all test stimuli are delivered by an audio recording, in the form of questions posed to the examinee in the target language of the test. Although this approach is reasonably flexible (questions on a variety of topics and at differing levels of sophistication can be presented) the oral question-answer technique was considered less well adapted to the efficient and objective elicitation of such aspects of overall proficiency as narrative discourse, supported option, and speech within role-playing situations than was a combined tape-and-booklet format which would allow the examinee to respond to pictorial and other printed material as well as to spoken questions. As one example, it was felt that past-tense narration on a specified topic would be much more readily elicited through a series of line drawings, whose "story" the examinee would be asked to tell. Audio-only stimuli would necessarily be complex and subject to only partial comprehension or misinterpretation by the examinee as to the production task involved.

3. *To the extent possible, the test should follow the normal progression of the live interview with respect to the types of questions posed and the level of linguistic sophistication required at any given point in the test.*

As is the typical sequence in the face-to-face interview, it was intended that the CST would begin with requests for simple factual information of an autobiographical nature. It would continue through narrative descriptions, narration in past time, responses requiring comparisons, supported opinion, etc., up through a final "situation" section. CTS probes, among other things, the examinee's ability to tailor the content and style of speech to the sociolinguistic requirements of a given interlocutor/communication situation combination.

4. *To the extent possible, questions throughout the test should be framed so as to represent genuine communicative tasks rather than artificial uses of the language.*

It was considered reasonably straightforward from a technical point of view, and of great benefit to the face validity of the overall test, to present each test question or series of questions within a "scenario" that would involve the examinee as a transmitter of genuine information to an identified interlocutor, rather than as a rote responder in an artificial linguistic setting. Instead of simply asking the examinee to "answer the following questions," the task was phrased as follows: "In the first part of the test,...you will have a conversation with a native speaker of Chinese. This person would like to find out some things about your background, activities, and interests." In a later part of the test, for a series of pictures depicting a burglary in progress and intended to elicit past time narration, the corresponding instructions were "In Taiwan, you have witnessed a nighttime breaking-and-entering episode. A policeman who has just arrived on the scene asks you to tell him what you observed. Based on the pictures shown, tell the policeman, in as much detail as possible, exactly what you saw happen." It was felt that this scenario approach informed the examinee, in a very straightforward and natural manner, of the identity and status of the interlocutor ("e.g., a policeman"), of the communicative task involved ("tell him what you observed"), of any relevant details about the task ("in as much detail as possible") and, for the more proficient examinee, of the speech style/discourse level at issue.

5. *Each question in the test should be answerable, to at least a limited extent, by examinees at the lowest end of the score range.*

This requirement was based on the anticipation that lower-level examinees would become highly frustrated by any questions to which they were at a complete loss to respond. To minimize this possibility, each question was carefully reviewed from the perspective of a typical "level 1" examinee, to insure that at least some rudimentary answer could be provided in all instances. For example, even in the final part of the "situation" section – which presented such potentially sophisticated challenges as "make appropriate remarks to your (Chinese hosts at a farewell reception), thanking them for their hospitality and expressing your hopes for continuing contact and further cultural exchange opportunities in the future" – a level 1 examinee would presumably be able to produce the Chinese equivalent of "thank you, we enjoy, we hope come again" or other similar responses.

6. *Wherever feasible, each question should be structured so as to provide latitude for elaborated responses on the part of the more proficient examinee.*

Although it was acknowledged that it would not be possible to provide a level 1 through 3+ range of "challenge" within every question in the test, an attempt was made to provide an opportunity for higher-level performance within as many questions as possible. For example, in a picture-based question in which the examinee is asked to tell a Chinese friend "what facilities are available in (an American recreation park) and, in general, what kinds of things there are to do there," the pictured scene includes – in addition to such basic items as a basketball court and swimming pool – people riding horseback, a sandbox, swings and slides, snack bar with cotton candy, and several other objects and activities that the level 3 or higher examinee might be expected to readily describe or paraphrase.

Test Content

Based on the overall test format/procedural criteria described above, a preliminary test form consisting of four basic sections was developed. A description of the content of each of these sections and further discussion of the measurement rationales involved is given below.

1. *Simulated Conversation*

In the first section of the test, the examinee listens to conversational questions spoken by a Chinese voice on the tape and responds appropriately in the target language. The questions posed are typical of polite conversation among relative strangers (which is usually the

implicit communicative situation in the early portions of the face-to-face interview), and include, for example, the idiomatic Chinese equivalents of:

Hello, how are you?
May I ask your name?
Can you tell me something about your family?
Are you working or studying?
That's interesting – can you tell me more about it?
Why at first did you want to study Chinese?
What do you enjoy doing in your spare time?
What kind of music do you like, and why?
What do you plan to do later today?
I hope we'll have a chance to talk again. Goodbye.

Several considerations in formulating questions for inclusion in this section may be noted. First, it was felt that each question should be applicable to the biographical experience of any person who might reasonably be expected to take the test. On this basis, a question such as "Can you tell me about your present studies?" would be inappropriate for any examinees not currently following an academic program. Second, it was necessary for each subsequent question to be a reasonable, or at least not illogical, continuation of any particular reply the examinee might have made to the preceding question. For example, to the initial "Hello, how are you?" question, a taped reply of "That's good . . ." (followed by another question) would not be an appropriate conversational follow-up for an examinee who had actually replied "I've got a splitting headache" or "I'm very nervous about taking this test." Third, the time provided for examinee response to any given question needed to be very carefully gauged to insure reasonably smooth and natural interplay of tape and examinee responses.

2. *Picture-Based Discourse*

This portion of the test consists of both single line drawings and series of topically related drawings, which serve as the stimuli for narrative descriptions or other communicative tasks on the examinee's part. For example, the first question in this section, focusing on the examinee's ability to give oral directions, consists of a bird's-eye view of a simplified route map, together with tape recorded English instructions as follows: "Imagine that you have invited a Chinese friend to come to dinner at your house, which is a few miles out of town. You need to give your friend directions over the telephone. These directions are shown by a dotted line on the map. You will have 15 seconds to study the map. Then, when you hear a tone signal, tell your friend how to drive from

point X at the bottom right of the map to your house, shown by an arrow at the bottom left."

As is the intent with all questions in the test, the situation incorporates a real-life communication task: giving directions on the phone. The particular linguistic content of the task also represents "genuine" communication of the examinee's part, in that it involves conveying initially unshared information to a specified interlocutor. As previously indicated, the attempt is made in the picture description section (and elsewhere throughout the test) to provide an opportunity for the lower-level examinee to make at least a minimal response: in this example, perhaps saying something like "Turn left, go straight. See school, turn left, go long time, house on left." At the same time, the drawing is sufficiently detailed to allow more proficient students to make appropriate mention of any of several additional features (e.g., a bridge, the indicated mileage beyond the bridge, a winding road through woods, pictured compass reference points) in giving their response.

The second, sequence-of-pictures portion of this section deals primarily with narration of habitual activities (e.g., examinee describes the "typical day" of an American office worker, based on a series of pictures provided); past time autobiographical narration (examinee describes to a doctor a recently experienced health problem); and hypothetical situations (examinee describes a sequence of steps that a friend will "probably have to go through" in renting an apartment). In all instances, depending on the proficiency of the examinee, the response can range from "survival" level discourse (e.g., "gets up, eats, goes to work, works hard, comes home five o'clock") to much more sophisticated narration, which in the case of the office worker scene, for example, is even open to relevant social commentary concerning "time pressure," "overwork," "traffic jams," and so forth.

3. *Responses to Printed Questions*

For each question in this section, the examinee reads a printed English question (which is also spoken aloud on the tape) and answers it in the pause provided. This section is intended to constitute an extension of the first, "conversation" part of the test into areas requiring longer discourse on the part of the examinee and to reflect such aspects of language use as making comparisons (e.g., "Which do you prefer, Chinese or American food, and why?"); stating opinions ("What, in your opinion, are some of the most important aspects of American life and culture that you would want a Chinese visitor to the United States to learn about or experience while in the U.S.?"); and stating and supporting positions on reasonably complex and/or abstract topics

("What do you consider the most serious political or economic problem facing the United States today, and on what do you base this opinion?"). Although it would have been possible simply to pose these types of questions orally in Chinese as part of the "simulated conversation" section, such an approach was decided against in view of the strong possibility that the idiomatic Chinese required to properly phrase the stimulus questions would pose a substantial problem of aural comprehension. In the absence of any repetition/clarification opportunity on the examinee's part, a mishearing or lack of full understanding of the aurally posed Chinese question could seriously hamper or prevent response, even in situations where the examinee's speaking proficiency per se was quite adequate to the production task involved.

4. *Situations*

The last section of the test was intended to parallel the variety of role-playing activities often included at the end of the face-to-face interview. Of all the question types in the test, the "situations" were considered to offer the most detailed and natural opportunity to assess the examinee's ability to select and use appropriate lexicon, register, and style of speech for a given interlocutor and communicative intent. Examinee instructions for this section provided the following information: "For each situation, you are asked to pretend that you are actually in the situation, and to respond in a linguistically appropriate manner . . . When you respond, be sure to use the type and style of language that a native Chinese speaker would use in this situation." The following descriptions, which were both printed in the text booklet and read aloud on tape, are exemplary of the types and range of communicative tasks involved.

> You have just entered a moderately priced restaurant in Beijing. When the waiter arrives, order a meal including a meat dish, a vegetable dish, cooked rice, and hot tea. Tell the waiter that you are in a hurry and ask if there would be any problem being served within the next ten minutes.
>
> You are standing in line at a train station ticket window in Nanking. There are two people ahead of you and your train is scheduled to leave in five minutes. Ask the people in front if they could possibly let you get ahead of them, explaining the reason for your request.
>
> You have arranged a trip to Shanghai through a Chinese travel agency in New York. When you arrive in Shanghai at

> the hotel that was suppose to have a room reserved for you, the clerk tells you there is no room in your name and, furthermore, that the hotel is completely booked for the evening. Irate, you go to the travel agency's local office and confront the agency representative. Inform him of your problem in detailed and appropriately indignant terms, and demand that suitable alternative lodging arrangements be made immediately.

For each situation, the physical setting; interlocutor; communicative task to be accomplished; and, by implication, the physical and emotional state of the examinee are indicated. Across the five situations in the test, the communicative requirements range from basic survival tasks (ordering a restaurant meal) to fairly sophisticated representational activities (making appropriate remarks to Chinese hosts in a ceremonial setting). The presumed interlocutors are also varied in social level and relationship to the examinee, and the presumed state of mind and physical status of the examinee range from neutral-but-slightly-hurried (restaurant scene) to tired and angry (hotel booking problem) to well-wishing and expansive (farewell reception). Although all of the situations are straightforward enough to be tackled in a very limited way by level 1 examinees, they are also of a nature that would pose a genuine challenge to higher-level students with regard to proper register, choice of lexicon, use of non-neutral intonation patterns or other affect-bearing phonological features, and so forth.

Clinical Administration and Analysis

At present, the preliminary form of the Chinese Speaking Test has been administered to a total of 33 undergraduate and graduate-level students of Chines at six institutions: Cornell University, Defense Language Institute (Monterey), University of Hawaii, University of Pennsylvania, Stanford University, and the Inter-University Program for Chinese Language Studies in Taipei. Student responses to the draft test materials, as well as information provided by written feedback questionnaires and debriefing sessions with the students, suggest that the testing procedure is functioning quite well overall, that is to say, is eliciting a wide range of student speaking performance in a straightforward and fairly natural manner. Again on only a clinical basis, the taped responses appear to be sufficiently extensive and differentiated across examinees to permit assigning ACTFL/ILR ratings (within in the intended test range) with a reasonable degree of confidence.

Although the results of this small-scale, clinically-oriented adminis-

tration have provided some indication that the assessment of performance on the basis of a proficiency-oriented semi-direct test of the type under development in the project will be generally congruent with the ACTFL/ILR rating scale, this will be much more rigorously investigated in a formal concurrent validity and scoring reliability study to be conducted in the fall of 1985, when all four test forms will be available. The intended design will involve administration of two forms of the tape-mediated test to each of 72 students, with pairing of forms and order of test administration arranged so as to balance the frequency of administration across tests and control for practice effects. The scoring reliability portion of the analysis will be based on the re-rating of a fifty-percent sample of the student response tapes by a member of the rating group other than the initial rater. Concurrent validity and reliability information will be sought both for the global level score and for individual "factor" ratings of pronunciation, listening comprehension, vocabulary, structure, and fluency.

Pending the detailed statistical results of the large-scale test administration, quite a bit of information relevant to test format, examinee instructions, and administration procedures has already been derived from the clinical administration. First, from the point of view of test format, the initial trials have suggested the desirability of providing ample time for any reasonable response to a given question, as opposed to a more parsiminious allocation of response time. Although the latter approach would economize on testing time[4] and would also make the between-question wait less onerous for examinees at the lower end of the proficiency range, the considerable frustration reported by more competent examinees at being "cut off" in several instances before they had responded to their own satisfaction militates in favor of providing a response time that is comfortably adequate for even a fairly elaborated response. For less proficient examinees, to minimize the potentially disconcerting effect of having to wait "speechless" for the next question, the test instructions will be revised to urge the examinee to say as much as is needed to fully accomplish the communicative task at issue, but not to be concerned if in some instances more response time is provided than seems necessary.

A second proposed modification based on the results of the clinical testing will be to reduce somewhat the amount of English associated with the test format and examinee directions. Although all examinee responses are in Chinese, and although the total test provides for a wide sampling of target language proficiency, the fact that both the printed and spoken test directions, as well as the printed and spoken stimuli for the "responses to printed questions" and "situations" sections, are in English tends to give the erroneous impression that relatively little Chinese is being dealt with by the examinee. Both to

minimize this impression and to somewhat shorten the total testing time, the test instructions will be revised judiciously to reduce the amount of English used, without jeopardizing the conveying of adequate information concerning the communicative task involved, the identity of the presumed interlocutor, and relevant affective considerations, all of which are considered vital to the proper operation of the test and to the validity of the elicited speech samples.

A third matter, which involves both practical and theoretical considerations, is the nature and extend of various types of "cultural load" in the test and the implications which this would be expected to have for examinee performance and interpretation of testing results. Across all of the questions in the clinical version of the test, the "simulated conversation" questions may be considered culturally neutral in that they represent the kinds of biographical and other types of "small-talk" questions that might be asked in virtually any geographic or cultural setting. The "picture-based discourse" section presents, for the most part, "American" scenes, including a western-style suburban home, train station, recreation park, an American worker's day-at-the-office, and steps in finding a U.S. apartment. (The two other picture series – a breaking-and-entering scene and a description of health problem – take place in a Chinese setting.) Items in the "printed questions" section are for the most part culturally neutral, although two – description of aspects of American life and culture of interest to a Chinese visitor; discussion of political/economic problems currently facing the U.S. – deal with American topics. All five "situation" tasks in the last section of the test have mainland China or Taiwanese settings.

Although there seems to be a reasonable balance of Western and Chinese settings within the test as a whole, the question still arises as to whether it is in some sense "unfair" to ask a student whose exposure to the language has never gone beyond the American classroom to deal in detail with Chinese settings and culturally determined linguistic aspects of these settings. On the other hand, for students who have been studying intensively abroad and have been immersed in an authentic Chinese environment, a task such as describing an American house in detail might be inordinately difficult by comparison to providing a similarly detailed description of housing accommodations in Taipei. The general question of cultural load and "fairness to examinees" is one which the test development group will continue to address, but the current orientation is to continue to insist fairly strongly on the concept of curriculum-free, generalized proficiency as the major touchstone for test content and interpretation, and on this basis to include in the test a reasonably balanced representation of the kinds of communicative situations and areas of discourse that would be expected to arise for an English-speaking learner of Chinese in the

course of his or her contacts with native Chinese speakers in both U.S. and foreign contexts.

By way of brief conclusion, development of the Chinese Speaking Test provides a substantial opportunity to attempt to reproduce, to the greatest extent possible, the content and manner of operation of the live proficiency interview within the technical parameters of a semi-direct format, and to compare the language-use and measurement results obtained through each of these procedures. A reasonably intensive and quite productive "dialogue" between the two types of instruments has already taken place in the course of development and trial administration of the semi-direct test. The subsequent large-scale administration of both instruments, together with the linguistic and statistical analysis of these results, should expand this dialogue appreciably further, to the potential benefit of both interlocutors.

References

Clark, J. L. D. (1979). Direct vs. semi-direct tests of speaking ability. In E. J. Brière & F. B. Hinofitis (Eds.), *Concepts in language testing: some recent studies*. Washington, DC: Teachers of English to Speakers of Other Languages.

Clark, J. L. D. & Swinton, S. S. (1979). *An exploration of speaking proficiency measures in the TOEFL context*. Princeton, NJ: Educational Testing Service.

Ervin, G. L. (Ed.). (1983). Pathways to change in language testing [Special Issue]. *Modern Language Journal, 67*(4).

Higgs, T. V. (Ed.). (1982). *Curriculum, competence, and the foreign language teacher*. Skokie, IL: National Textbook.

Higgs, T. V. (Ed.). (1984). *Teaching for proficiency: the organizing principle*. Lincolnwood, IL: National Textbook.

James, C. J. (Ed.). (1985). *Foreign language proficiency in the classroom and beyond*. Lincolnwood, IL: National Textbook.

Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: a historical perspective. In T. V. Higgs (Ed.), *Teaching for proficiency, the organizing principle* (pp. 11-42). Skokie, IL: National Textbook.

Lowe, P., Jr., & Clifford, R. T. (1980). Developing an indirect measure of overall oral proficiency. In J. R. Frith (Ed.), *Measuring spoken language proficiency* (pp. 31-39). Washington, DC: Georgetown University Press.

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (Ed.), *Direct testing of speaking proficiency: theory and application* (pp. 1-12). Princeton, NJ: Educational Testing Service.

Footnotes

[1]For a description of the development background and measurement rationale of the FSI interview technique and scoring scale, see Sollenberger (1978).

[2]This project, directed by the writer and co-directed by Dr. Ying-che Li (University of Hawaii), is formally titled "Development, Validation, and Dissemination of a Proficiency-Based Test of Speaking Ability in Chinese and an Associated Assessment Model for Other Less Commonly Taught Languages," and is supported by the U.S. Department of Education under grant No. G008402258, Center for International Research and Studies. Members of the project working committee include, in addition to the co-directors, Dr. Albert Dien (Stanford University), Dr. Shang-hsien Ho (University of Hawaii), Dr. Timothy Light (Ohio State University), Dr. Eugene Liu (University of Pennsylvania), Dr. Pardee Lowe (Defense Intelligence College), and Dr. A. Ronald Walton (University of Maryland).

[3]ILR level 3+ is somewhat above the lower bound of the ACTFL "Superior" rating, which coincides with any level 3 or higher performance on the ILR scale.

[4]Tape-running time for the preliminary version, including examinee directions, spoken stimulus material, and response pauses, is approximately 35 minutes. Examinees responding fully to all questions would speak for a total of about 16 minutes.

Using Technology to Assess Second Language Proficiency Through Cloze-Elide Tests

Winton H. Manning
Educational Testing Service, Princeton, New Jersey

147

Abstract

The aims of the projects described in this paper are to develop and evaluate a new approach to the assessment of English language skills of ESL students. The approach makes use of two technologies: (a) an innovative use of optical scanning equipment in scoring cloze-elide test exercises, thereby providing a rapid, relatively inexpensive and objective measure of language proficiency; and, (b) the use of a microcomputer as a testing medium, employing an adaptation of this same cloze-type measure. Broadly speaking, the projects have two goals:

1. *Developmental.* To devise a cloze assessment procedure that might be useful in the sorts of testing applications ETS often encounters — large testing programs that usually utilize standardized tests that are amenable to rapid, objective and cost-effective scoring.

2. *Research.* To investigate assessment methods that offer greater promise of being grounded in psycholinguistic theory than do traditional multiple-choice, reading comprehension tests.

Using Technology to Assess Second Language Proficiency Through Cloze-Elide Tests

Cloze-Testing

Over 30 years have passed since the publication by Wilson Taylor (1953) of his paper describing the "cloze" procedure. In the 3 decades since that publication cloze testing has increased enormously, and research on cloze methodology has grown apace (Oller, 1979, 1983). Recent estimates from Educational Resources Information Center (ERIC) suggest that in the past 10 years over 750 articles and research publications have appeared that are concerned with one or more aspect of cloze testing methodology.

It is now generally recognized that the cloze testing procedure is a good technique for the assessment of second language proficiency. Among the several features that recommend its use are ease of test development (Rye, 1983) and relatively high correlations with other, more traditional tests of reading comprehension (Anderson, 1976; Hinofotis, 1976; Weaver & Kingston, 1963).

The cloze procedure has attracted the interest of many language teachers and researchers. In part this interest has grown because of changes that have occurred in the field of linguistics, and the impact of linguistics upon language testing. Some authors have observed that those interested in language testing have become more concerned with the construct validity of measures of language proficiency, rather than the statistical assumptions and properties of various kinds of tests. These developments are primarily due to rethinking what it means to understand a language, and to the development of new methods of language teaching. Cziko (1980) has said that: "The most important change in language testing has been the introduction and use of *integrative* tests of language proficiency (e.g. cloze and dictation tests) in contrast to the more traditional though still popular *discrete-point* tests (e.g. vocabulary and grammar tests)" (p. 27).

Whether or nor one fully agrees with this assessment, it is probably correct that so-called integrative tests of language skill offer real promise as useful measures although, heretofore, their attractiveness has been diminished for use in large scale testing programs such as those often carried out by ETS because of the near impossibility of

scoring such tests rapidly and efficiently, by machine.

A factor that has limited more widespread use of cloze testing is the necessity of scoring traditional cloze tests by hand, using either "exact" word or "acceptable" word keys, or occasionally more complex rules for evaluation responses of examinees. For teacher-made tests this labor-intensive scoring requirement is not ordinarily a heavy burden, although it may be an irksome and inefficient use of teacher time. However, the need for large scale use of cloze tests that are standardized and that can provide normative data has inevitably led to the development of multiple-choice (m-c) cloze tests (Hinofotis, 1976; Hinofotis & Snow, 1980; Jonz, 1976; Baldauf & Propst, 1979). Typically, a list of several possible words is provided for each missing word, and the student must then choose the "best" word to be inserted into each blank space in the text. Responses may be obtained through use of a conventional, m-c format, and machine-scorable answer sheets. The m-c cloze methodology has been found useful in a number of studies (e.g., Fitzgerald and Fitzgerald, 1978) and the TOEFL program is currently sponsoring a study of m-c cloze procedure in assessing TOEFL candidates in a project now underway (Hale, 1983).

Some critics have pointed out that the m-c format may fundamentally change the cloze task from one of *producing* the correct word to one of *recognizing* it (Valmont, 1983). For purposes of assessment, as compared with instruction, such distinctions may not be of crucial importance. The cloze-elide exercises that we are developing do depend heavily upon *recognition* but the task appears to be novel and appealing. Performance on the task requires the processing of textual material not only in a way that is much less structured than in m-c cloze tests but also less structured than a traditional cloze test.

Cloze-Edit Tests as Examples of Cloze-Elide Testing

As stated earlier, in this research we are exploring a testing format that differs from the traditional cloze testing procedures, but which also seems especially appropriate for application in large-scale testing programs. This technique is the *Cloze-Edit* procedure first described by Davies (1975), and empirically studied by Bowen (1978). However, in our work we have modified this procedure to permit its scoring by optical scanning equipment and in this paper we shall refer to the method as cloze-elide testing. The cloze-elide technique requires the test taker to strike out irrelevant or incorrect words that have been inserted into a prose passage. As a generic term for this general approach the term "*cloze-elide testing*" has been proposed by the author since there are many different variations that are possible, and since the procedure whereby such test exercises are rendered scorable by optical scanning machines is a new development.[1]

Although the cloze-elide procedure bears a sufficient resemblance to traditional cloze testing to merit that designation, it is important to note that there are differences, and that these differences may turn out to be interesting or even important. In a traditional cloze test, words are deleted from the text according to some rule. From the point of view of the respondent, however, the eye is drawn to the particular points where gaps occur. The task of the subject requires that the actual missing word (or one that is acceptable) be inferred from the meaning of the remaining prose, utilizing what some have called the "grammar of expectancy" (Oller, 1983). There is, nevertheless, a sharply defined perceptual boundary between what is given (the mutilated text) and what must be supplied (the missing word).

In cloze-elide tests, words have been *inserted* into the text, according to some prescribed format. The respondent perceives a printed page, superficially like any other. The inserted, extraneous, stimuli serve to "break-up" the meaning—to interfere with the cognitive processing of the language. In our pilot work we have interrogated students, asking them to describe or introspect on their own cognitive processes as they work through a cloze-elide exercise. It seems that students see the sentence or phrase as the stimulus, in which, because of the extraneous words, the *meaning* is broken. When the inserted words are marked out the "meaning comes together" in almost a physical sense. Students speak of this in terms of a self-reinforcing, positive gratification — the achievement of meaning which emerges as the extraneous material is deleted. This is in contrast, we believe, to the reports of frustration and lack of fulfillment that students often report in describing their experience in completing a traditional (fill in the missing word) cloze exercise. Of course, these observations are at this point only preliminary hypotheses and need further study. When one considers a cloze-elide exercise it is evident that, literally, any word *could* be extraneous. There are no superficial cues that mark off the boundaries between prose that is part of the text, and those words that must be eliminated from the text in order for the meaning to be reclaimed from the noise that has been interjected. I contend that this response, too, is an example of an integrative task in which the "grammar of expectancy" plays a significant role. What is different is that the surface cues are more ambiguous in cloze-elide tests than in traditional cloze tests, or in m-c cloze tests. Whether this change is an important difference or not, in terms of assessing underlying language proficiency, is unknown at this time. It is possible, however, that it may be important, and that cloze-elide tests may thereby tap into underlying skills in language that are somewhat different than other cloze-type tests. Of course, this analysis is also at this point a speculative one, and one objective of our work is to shed some light on this issue.

An example of a cloze-elide exercise may be helpful:

> Bats are among the world's most diverse and train fascinating mammals. Nearly a thousand make kinds comprise approximately a fourth of all mammal species, and they poetry inhabit nearly every part of earth. They discover range in size from a bumble physics bee-sized bat that is quarter the world's smallest mammal to slave giant flying foxes with six-foot wingspans. Although bats are, for their sun size, the world's longest lived mammals tooth (more than 30 years), their highly specialized festival life styles and unusually boat slow reproductive rates (usually only one hair young per year) make adaptation to habitat avenue changes extremely difficult. Furthermore, bats form astronaut the largest and most vulnerable helmet colonies of any warm-blooded flame animal.

Such exercises are relatively easy to create, although care must be exercised to assure that the extraneous words are, indeed, lexically or syntactically incorrect. We are also interested in studies that would examine the interactions of inserted words, at varying locations, in relation to the semantic structure of the sentence, and in relation to the parts of speech preceding and following the inserted words, but these topics are not addressed in this paper.

The procedure for producing cloze-edit exercises by altering reading passages is described later but the basic approach to test administration is as follows: specially prepared *cloze-elide* passages are printed on optically scannable answer sheets and the examinees are required to eliminate the incorrect, inserted words from the text by drawing a line through the words with a pencil.

The altered test is printed on the answer sheet/test booklet in non-reflective ink, so that only the graphite deposited by the students' pencil stroke may be read by the scanner. This procedure has been developed by the author at ETS, and a patent application has been approved by the U.S. Patent Office to protect ownership of this invention and certain related techniques by which cloze-elide passages could be generated and scored by means of a computer.

Although the general approach of cloze-elide testing appears to have wide applicability, its use in ESL testing seems especially appropriate, since the method of cloze-elide testing, if successful, would permit inclusion of test exercises that are machine-scorable, and while not in m-c format, are nevertheless conceptually similar to well known cloze techniques.

Prior Research on Cloze-Elide (Cloze-Edit) Exercises

In 1978, J. D. Bowen (1978) briefly described a variant of the cloze testing procedure as an indirect measure of second language proficiency. It appears that independently Davies has been experimenting in Great Britain with this technique which he described as "interposing English distractors randomly in a running text" (Davies, 1975, p. 124). In our procedure irrelevant words are inserted at random points in an English or foreign language text, and the task of the examinee is to "edit" or elide the words that have been inserted. Bowen (1978) reported substantial validity for cloze-edit test scores with subtests of the Michigan Test of English Language Proficiency, the Michigan Test of Aural Comprehension and a test of written composition. He concluded that the editing test was "...reliable, practical and valid." (Bowen, 1978, p. 15). Mullen (1979) later undertook a similar study in which she also developed "cloze edit" passages and evaluated their effectiveness in the assessment of English language proficiency of foreign students enrolled in ESL classes at the University of Louisville. She reported high correlations with traditional cloze procedures, strong but moderate correlations with some subtests of TOEFL, and high concurrent validities with both a graded essay and with a measure of language comprehension obtained through an oral interview. Mullen's results are interesting and suggested to us that cloze-elide tasks offered considerable promise in the assessment of proficiency in a second language.

Recently, work has been carried out at the University of Washington concerned also with evaluating the usefulness of cloze-elide tests. Under the direction of Brita Butler-Wall, samples of ESL students at the University of Washington were tested, using specially constructed passages of varying difficulty and style (Butler-Wall, 1984, personal communication). Butler-Wall has tentatively concluded that "the editing task is a promising tool for ESL administrators and teachers".

The available evidence, though limited, suggested to us that the cloze-edit exercise has substantial promise as an alternative to traditional cloze-tests. However, as Bowen observed, in its standard form it is very tedious and difficult to score, since errors of "omission" and "commission" are possible. As noted earlier, recasting the edit task into the scannable cloze-elide format avoids the difficulty of scoring that Bowen has deplored. A major objective of our research is to provide a more thorough evaluation of the cloze-elide measure than earlier research has provided, with particular attention being given to its potential use in services provided by the TOEFL program, or in foreign language testing.

Scoring Cloze-Elide Tests

Inspection of the example of a cloze-elide exercise will make clear the fact that a student can make two kinds of errors when completing the exercise. These are:

1. *Errors of omission* (O) occur when an inserted word is *not* crossed out. The maximum number of such words is the number of extraneous words that have been inserted into the passage.
2. *Errors of commission (C) occur when the subject crosses out a word that is deemed integral to the test (i.e. a word that has not* been arbitrarily inserted). It is clear that C errors can range from deletion of words essential to the preservation of meaningful syntax or mere reflections of a stylistic preference by the student.

Early research on cloze-edit exercises handled the problems of these two error scores, which are partially *dependent* rather than independent variables, by treating them as two separate scores.

Our approach has been to calculate a "formula" score for cloze-elide (c-e) tests, which corrects for random guessing, and combines the two scores as follows:

$$S_i = R_E - E\left(\frac{A}{A+T}\right)$$

where:

S_i = formula score for a student on a CE exercise;

R_E = number of extraneous words correctly elided;

E = total number of words elided;

A = number of extraneous words added to the reading passage;

$A+T$ = total words in the reading passage.

This expression has some desirable properties: (a) if the elisions are done randomly then the expected value of S is zero; (b) but if all words are elided or no words are elided then $S_i = 0$, also.

The effect of this formula score is to combine the two separate error scores — those of commission (C) and omission (O) — in a model that logically synthesizes them, and avoids the problems inherent in the partial linear dependency of C and O errors. Our models for treating this problem have been explored, but for practical purposes since they are so highly intercorrelated (greater than .98), they are not described

in this paper. The alternative, of assessing performance on a single task by two separate but mathematically dependent error scores, is, we believe, much less desirable.

Finally, if a cloze-elide test has several items, then a raw score is of the form:

$$\sum_{i=1}^{n} w_i S_i$$

where the weight w_i is chosen to reflect how much weight the tester wishes to put on passages of different lengths.

Two Research and Development Projects

At the present time two studies are underway to evaluate the construct validity and practical usefulness of cloze-elide testing. The first of these is a study supported by the TOEFL policy council. This study will look at the relationships among cloze-elide tests, other cloze measures, writing ability, and TOEFL scores of students more or less representative of the TOEFL population. Instruments are only now being developed, and we have not yet collected data. For that reason I will not discuss the TOEFL study further at this time. The second study which concerns assessment of language proficiency of younger students, at the elementary and secondary levels, is further along. It is, for convenience, termed the Elgin Study.

The Elgin Study

In 1986 the National Assessment of Educational Progress (NAEP) will carry out a study of English language proficiency of students for whom English is a second language. This study will afford an unusual opportunity to gain an understanding of the language proficiency of this large and growing population of school age children. The project I will briefly describe is part of the preparatory activity for this later NAEP assessment. Our primary goal is the concurrent validation of several potential approaches to evaluate language proficiency, and the subsequent choice of the most effective measures that we can devise, for possible inclusion in the 1986 NAEP language survey.

Several constraints operate to limit or influence the choice of instrumentation for the operational NAEP language survey. These are:

1. Only group administered paper/pencil instruments can be used.
2. The instructions must be self-explanatory so that students can

undertake the test exercises without further teacher assistance or explanation.

3. Language proficiency testing is constrained to a total of 6 minutes, including silent reading of instructions.

A further objective is that:

4. It would be desirable to incorporate into the NAEP language probe measures of proficiency that are innovative and positive in their influence on teaching and assessment practices for language minority students.

During the winter and spring of 1985 a project was mounted in the Elgin, Illinois schools which had as its purpose the field evaluation of several promising candidate exercises for assessing language proficiency, at the three age/grade levels.

This evaluation has now gathered evidence concerning the validity of these "exercises" for more extensive, widely used, measures of reading, writing and other language skills. The study also permitted assessment of the reliability of the candidate exercises, the degree to which instructions are self administering (especially for elementary school children), and the amenability of the candidate measures to the severe 6-minute time restraint.

Variable Descriptions

The 10 measures used in the Elgin Study were as follows:

1. *Writing (Graded Essay).* A 25 minute composition on a free-writing topic. Essays were rated by two experienced readers, who were knowledgeable about language minority children. The rating scale ranged from 1 to 6.

2. *Reading Comprehension (NAEP).* A number of reading exercises of the kind used in the NAEP program. These consist of a short reading passage followed by several m-c questions based on the passage. There were a total of 18 items at grade 3-4 and 22 items at grades 7-8 and 10-12.

3. *Reading Comprehension (MECCA).* A subsection of the Elgin Schools' MECCA Test, consisting of a modified cloze procedure. It consists of a reading passage with 25 deletions. Each item has 4 m-c options. Grade level appropriateness for the passage was done by applying the Fry readability formula.

4. *Receptive Skills (Listening).* A subsection of the MECCA test, this is an auditory measure of functional language use. It consists of 10 items based on aspects of the school experience (e.g. school procedures, classroom practices).

5. *Expressive Skills (Speaking).* A subsection of the MECCA test, this is a speaking measure of functional language use. It consists of 10 items based on the school experience (e.g. naming classroom objects, report on classroom rules). Students responses are obtained individually and evaluated according to prescribed standards.

6. *Language Total (Sum 3, 4, 5).* Overall score for the MECCA test. There is a total of 45 items: 25 reading comprehension, 10 receptive skills and 10 for expressive skill.

7. *Picture Vocabulary.* Used to measure English vocabulary knowledge of 3-4 graders only, this test was originally developed for native Spanish speakers. It consists of 27 items. Examples of items are location words, simple action verbs, synonyms and use of words in a sentence context.

8. *Standard Cloze (Exact).* These were reading passages in which every 10th word was deleted. Exact word replacement scoring method was used in this instance. Three reading passages of varying degree of difficulty were used with a total of 45 deletions.

9. *Standard Cloze (Approximate).* Same as above, but the "acceptable" word replacement scoring method was used, according to a set of prescribed rules.

10. *Reading (Multiple-Choice Cloze).* Target words were selected form the text, and two incorrect words were paired with the target word to form a 3-option item. One distractor was in the same word class; the second distractor was from a different class. Three different reading passages of ordered degree of difficulty were used. There was a total of 45 items.

Self-Rating of Students: For variables 11-14 students chose from a student questionnaire the answer that best described their background characteristics. Variable 15 – Overall Self-Rating – was a global self-assessment of English proficiency on a 6 point scale.

Teacher Appraisals: In variables 16-19 teachers rated 3-4 graders on receptive (listening) and experience (speaking) levels of proficiency in

the informal (playground) and formal (classroom) contexts. The scale was a 4 point scale. Variable 20 is a global rating of English proficiency on a 4 point scale.

Variables 21-25 rate the students achievement in the different subject matter content areas and also in a global assessment of school achievement. Each student was rated in comparison to the rest of his/her classroom.

Variable 26 is a global rating of cognitive ability independent of language proficiency and achievement.

Years of Residence Variable 27 – years of residence in the U.S. – is self-explanatory. (The variability for Laotian children was severely restricted, thereby rendering these correlations less meaningful.)

Results in the Elgin Study

Table 1 displays only a small portion of the data being analyzed in the Elgin Study, namely the validity of the cloze-elide tests for the criteria described in the Variable Descriptions section.

The Laotian sample is quite small, but is included, even so, for reasons of completeness and because the relationships with the test variables (1 through 10) are for the most part consistent with the results found in the larger Hispanic samples.

It seems evident from inspection of these validity coefficients that, for these ESL students the cloze-elide (CE) measure (a) is very strongly correlated with other cloze measures; (b) is well correlated with measures of writing and reading; (c) shows moderate to strong validity for listening and speaking; (d) displays strong correlations with a composite variable of language proficiency; (e) is substantially related to student self-ratings, though more for Hispanic students than in the small Laotian sample; (f) is substantially related to teacher appraisals of student's language skills; and (g) displays surprising (to us) good validity for English language proficiency of native English speakers.

Since data analyses are only partially completed, the data displayed in Table 2 should be understood as preliminary, and is included primarily because it constitutes initial evidence of the promise offered, we believe, by cloze-elide tests of language proficiency. In Winter, 1985, a more complete report on this project will be prepared, containing more extensive analyses.

In summary, we believe cloze-elide tests offer considerable promise as valid, indirect measures of language proficiency and that their adaptability to optical scanning offers potentially positive advantages for large scale testing applications where ease and cost-efficiency of scoring are important considerations.

Table 1
Validity of Cloze-Elide Tests of Proficiency in English for Hispanic (H), Laotian (L) and Native English Speaking (N) Students at Three Age/Grade Levels

Criterion Measures	Correlations with Cloze-Elide Tests Elementary 3/4		Middle 7/8			High 10/11/12		
Tests	H	N	H	L	N	H	L	N
1. Writing (Graded Essay)	65	34	57	76	40	67	53	41
2. Reading Comprehension	75	77	80	64	60	81	68	63
3. Reading (M-C Cloze)	83	51	73	83	40	83	74	74
4. Receptive Skills (Listening)	35	28	34	40	37	62	27	38
5. Expressive Skills (Speaking)	40	22	63	65	33	64	35	40
6. Language Total (Sum of 3, 4, 5)	77	54	78	82	52	85	66	73
7. Picture Vocabulary	63	69	—	—	—	—	—	—
8. Standard Cloze (Exact)	76	66	77	78	60	77	65	63
9. Standard Cloze (Approx.)	79	68	80	84	67	79	71	64
10. Maze (M-C Cloze)	78	71	87	87	80	89	75	73
Self Ratings of Students								
11. Understanding English	—	—	36	63	10	48	15	12
12. Speaking Skills	—	—	53	39	14	64	20	07
13. Reading Skills	—	—	56	37	16	62	36	28
14. Writing Skills	—	—	48	38	14	60	23	12
15. Overall Self Rating	—	—	58	46	46	50	25	35
Teacher Appraisals								
16. Informal Receptive	56	36	—	—	—	—	—	—
17. Formal Receptive	45	37	—	—	—	—	—	—
18. Informal Expressive	61	25	—	—	—	—	—	—
19. Formal Expressive	53	35	—	—	—	—	—	—
20. Overall English Proficiency	62	55	—	—	—	—	—	—
21. Science Achievement	42	47	—	—	—	—	—	—
22. Social Studies Achievement	37	51	—	—	—	—	—	—
23. Language Arts Achievement	62	52	—	—	—	—	—	—
24. Reading Achievement	55	57	—	—	—	—	—	—
25. Overall Achievement	59	49	—	—	—	—	—	—
26. Overall Cognitive Ability	21	49	—	—	—	—	—	—
Other								
27. Years Resident in U.S.	—	—	56	46	—	57	27	—

Sample sizes on which the data in Table 1 were calculated are presented in Table 2.

Table 2
Sample Sizes in Elgin Study

	Elementary Grades 3-4	Middle Grades 7-8	High Grades 10-11
Hispanic	114	113	144
Laotian	0	41	58
Native English Speakers	134	128	98
TOTAL	248	282	300

Microcomputer Based Cloze-Elide Tests

The rapidly growing availability of microcomputers in secondary schools and colleges offers attractive opportunities to develop new, different, and possibly more effective measures of reading comprehension in ESL situations or for students of foreign languages. It is important that research and development be undertaken that creatively utilizes this burgeoning new resource for improving teaching and assessment of language competency. Measurement techniques that take advantage of the flexibility and interactive capabilities of the microcomputer are important means of improving instruction and assessment for language teaching in our schools and colleges. Often in the past when computers have been used for test administration, conventional paper-and-pencil test procedures, using m-c questions, have been simply transferred to a terminal screen. The advantage of this use of computer technology lies mainly in the ability to provide immediate feed-back of results to the examinee or in shortening test administration time by using adaptive testing approaches. Often there is an insufficient justification for the time, effort, and resources that must be invested to produce the software for this kind of computer based test. In this project, we are undertaking an ambitious effort to develop cloze-elide exercises that appear to be especially appropriate to the processing capabilities of the microcomputer, that offer promise of being intrinsically interesting to students, and that would not only encourage the preparation of such tests by language teachers for classroom use but would also lend themselves to development of normed, standardized versions of the same type of tests as described earlier.

A Computer-Based Cloze-Elide System

Cloze-elide tests, we believe, especially lend themselves to the presentation of language assessment by means of a microcomputer as well as in paper and pencil formats. The basic approach to testing is as follows. Specially prepared cloze-elide passages are displayed on a microcomputer screen and the examinee is required to eliminate the incorrect, inserted words by deleting them from the text. Four technologies have been under consideration to permit students to respond speedily and easily.

With the decrease in the cost of microcomputers and peripherals and the increase in the reliability of these devices, alternative technologies for interfacing with microcomputers have been revolutionized.

These technologies are intelligent, simple alternatives that will provide high quality cursor control for the user. In addition, these technologies can interface with non-computer sources, such as video recorders, video-disc players and live camera with the proper computer programs. Popular high-level micro languages can support these devices.

The following is a brief review of devices and alternative technologies for interfacing with microcomputers that we have considered in developing a micro version of cloze-elide tests.

1. *Koala Pad Touch Tablet:* The Koala Pad touch tablet is a state-of-the-art digitized device that converts pressure resulting from the movement of a finger or stylus pen into electronic signals for controlling microcomputers at specific locations on the computer display screen. The display screen coordinates are available to programs resident in the microcomputer memory.

2. *Light Pens:* Light pens are more analogous to working with a normal pen, that is, the eye and hand work together on the screen. The microcomputer sends out signals to the CRT (Cathode Ray Tube) or TV which then displays the message on the screen. The screen is a collection of tiny dots or pixels which can be individually colored. For example, the Apple computer uses a square that is 280 pixels across and 192 pixels down. When the light pen is pointing to the screen, it "sees" a grid of pixels (280 x 192 or 53,760 dots). The screen display coordinates can be identified as a response from the user and are available to programs. Joysticks and game paddles utilize the same coordinates in the microcomputer memory.

3. *Mouse:* The mouse, named for its physical resemblance of the real thing, reports relative changes in cursor location. The mouse op-

erates on a flat surface by sending signals on its horizontal and vertical movements to the computer to update the position on the screen. A computer program can decode the signal from the mouse as a selection response from a user. The mouse is fast and accurate, and allows the user's hand to stay in the same position relative to the mouse at all times.

4. *Hewlett Packard (HP) 150 Touchscreen computer:* The HP 150 utilizes a 9 inch touchscreen — a grid of light-emitting diodes and photo diodes. There are 24 holes in each side of the screen and 40 holes in both the top and the bottom. This provides touch sensitivity for each row of the touchscreen display and for each unit of two columns. The system recognizes a touch when an object, such as your finger, breaks the light beams. The system shows its recognition by displaying the object in inverse video. To execute the selection or object, the user simply withdraws his finger and the system acts on the selection.

The advantages of direct screen devices, such as light pens and the HP 150 touchscreen, are that they turn the hand or finger into a conduit between the user and the computer. Direct screen devices are the easiest input device to learn and operate, but possibly tiring. The disadvantages of indirect screen devices, such as the mouse and Koala Pad, are that they require eye hand coordination. Also, the surface of the Koala Pad is much smaller in scale than the screen display.

We have recently completed the development of a prototype cloze-elide test, using the IBM-PC and light pen technologies. This strategy rests upon the belief that a simple, direct, and rapid response mode will facilitate student motivation and enable the text processing task to be completed in a real time frame that is more compatible with the act of reading (and editing) a text than would be keyboard entry of missing words. Unless the student is a touch typist — especially unlikely with children or foreign students — keyboard entry requires the examinee to gaze alternately at the screen and the keyboard. We feel this is an undesirable requirement for a language test.

Unfortunately, we have also found that the relatively inexpensive light pens we have used are not as reliable as we would like! For adults the problems are not serious but there might be a problem for children. As a consequence we are looking now at other light pens, and also considering moving to a mouse controlled cursor as the control mechanism.

Nevertheless, we judge the prototype to be a successful demonstration of the ease and suitability of the cloze elide exercise for administration and scoring by means of a microcomputer. Our hope is to press

forward with this development to the area of foreign language assessment as well.

Concluding Observations

As stated at the outset we are deeply committed to a program of developmental research aimed at (a) improving our understanding of integrative, cloze-type measures of language proficiency; (b) developing practical and useful assessment instruments for large scale, standardized test programs, as well as classroom use; (c) exploring ways in which optical scanning and microcomputer technologies can provide vehicles not only for improved delivery of tests, but also for improved, more effective measurement instruments; and (d) attempting to extend our understanding of the cognitive processes that underly performance on cloze-elide tests. This work has only just begun, but the results so far are encouraging.

References

Anderson, J. (1976). *Psycholinguistic experiments in foreign language testing*. St. Lucia, Queensland: University of Queensland Press.

Baldauf, R., & Propst, I. (1979). Matching and multiple-choice cloze tests. *Journal of Educational Research, 72*(6), 321-326.

Bowen, J. D. (1978). The identification of irrelevant lexical distraction: an editing task. *TESL Reporter, 12*(1), 1-3, 14-16.

Cziko, G.A. (1980). Psychometric and edumetric approaches to language testing: implications and applications. *Applied Linguistics, 2*(1), 27-44.

Davies, A. (1975). Two tests of speeded reading. In R. Jones & B. Spolsky (Eds.), *Testing Language Proficiency* (pp. 119-127). Arlington, VA: Center for Applied Linguistics.

Fitzgerald, T., & Fitzgerald, E. (1978). A cross-cultural study of three measures of comprehension at the primary and intermediate levels. *Educational Research Quarterly, 3*(2), 84-92.

Guthrie, J. T. (1973). Reading comprehension and syntactic responses in good and poor readers. *Journal of Educational Psychology, 65*(3), 294-299.

Hale, G. (1983). *The role of cloze items in the TOEFL*. Proposal submitted to TOEFL Research Committee. Princeton, NJ: Educational Testing Service.

Hinofotis, F. (1976). *An investigation of the concurrent validity of cloze testing as a measure of overall proficiency in English as a second language*. Unpublished PhD dissertation, Carbondale, IL: Southern Illinois University.

Hinofotis, F., & Snow, B. (1980). An alternative cloze testing procedure: multiple-choice format. In J. W. Oller & K. Perkins (Eds.), *Research in Language Testing* (pp. 129-133). Rowley, MA: Newbury House.

Jonz, J. (1976). Improving on the basic egg: the m-e cloze. *Language Learning, 26*, 255-265.

Mullen, K. (1979). An alternative to the cloze test. In C. Yorio , K. Perkins & J. Schachter (Eds.), *On TESOL '79: the learner in focus* (pp. 187-192). Washington, DC: Teachers of English to Speakers of Other Languages.

Oller, J. S., Jr., (Ed.). (1983). *Issues in language testing research*. Rowley, MA: Newbury House.

Oller, J. W., Jr. (1979). *Language tests at school: a pragmatic approach*. London: Longman.

Rye, J. (1982). *Cloze procedure and the teaching of reading*. London & Exeter, NH: Heinemann Educational Books.

Taylor, W. (1953). Cloze procedures: a new tool for measuring readability. *Journalism Quarterly, 30*, 415-453.

Valmont, W. (1983). Cloze and maze instructional techniques: differences and definitions. *Reading Psychology, 4*(2), 163-167.

Weaver, W., & Kingston, A. (1963). A factor analysis of the cloze procedure and other measures of reading and language ability. *Journal of Communications, 13*(4), 252-261.

Footnotes

[1]For example, it is possible to prepare machine scorable cloze-elide tests of the "maze" exercise described by Guthrie (1973), and to utilize the technique in tests of grammar, in teaching poetry and in other applications. The technique whereby tests that are developed by inserting words into a text according to a particular format but scored through optical scanning equipment or by microcomputers has been called "cloze-elide testing" by the author. The special answer sheet used for cloze-elide testing is the subject of a patent application, which has recently been granted by the U.S. Patent Office.

Using the Writer's Workbench In Composition Teaching and Testing

Joy Reid
Colorado State University

Abstract

During the past several years, the series of computer text-analysis programs called the Writer's Workbench (WWB) has been used in composition teaching and testing at Colorado State University. Research with native speakers of English has shown that students using the WWB significantly improve some writing skills, and that the quantitative measures of the WWB can identify some characteristics of holistically scored quality writing. This paper describes a year-long research project funded by the Exxon Education Foundation that focused on the use of the WWB in teaching and testing English as a Second Language (ESL) student writers. Results of the study indicate that ESL students who used the WWB significantly improved their editing and writing skills, that several WWB programs are potentially useful in evaluating ESL student essays, and that the two types of writing studied (description/interpretation of a chart and explanation/discussion of contrasting ideas) differ quantitatively.

Using the Writer's Workbench In Composition Teaching and Testing

In the fall of 1981, freshmen composition students at Colorado State University (CSU) began using word processing and text-analysis software programs. The Department of English Computer Laboratory initially consisted of an Onyx 16-bit microcomputer, with word processing software from Interactive Systems Corporation, and two Perkins-Elmer 1251-I terminals. Through a research exchange with Bell Laboratories, CSU received a computerized textual analysis package consisting of 32 programs called the Writer's Workbench (WWB). This text-analysis software was developed (Cherry, 1982; Cherry, Fox, Frase, Keenan & MacDonald, 1983; MacDonald, 1980; MacDonald, Frase, Gingrich & Keenan, 1982) to assist Bell Labs' technical writers (for a brief description of the WWB programs used as CSU, see Appendix A). Since the fall of 1981, half a dozen CSU faculty members have researched the feasibility, the effectiveness, and the potential benefits of using the WWB text-analysis programs in composition teaching and testing. The Exxon Education Foundation funded a study of the effects of the WWB on the composition skills of English as a Second Language (ESL) international student writers.[1] This paper reports briefly on several WWB studies completed in the past 3 years and discusses the investigation of non-native speaker writing and the use of the Writer's Workbench.

Background

The WWB has three distinct features. First, the programs interact with and analyze each student's own prose. For example, the number of "to be" verbs are counted in each student essay (FINDBE), the vague and abstract words in each essay are listed when these words exceed a set percentage of the text (VAGUENESS and ABSTRACT), and the percentage of short and long sentences (based on the student's average sentence) is identified (PROSE/Sentence Variation).[2] Second, the WWB offers suggestions for improving weaknesses identified in the prose. For instance, when DICTION flags a potential error, SUGGEST follows with one or more options for strengthening. Third, the WWB programs are programmable: virtually every program can be modified for the specific needs of the students. The PROSE program,

for example, compares a student's or professional's writing sample with selected papers typical of that level of writing. Different "standards" for different writers allow remedial students, for example, to compare their average sentence length with that of other basic student writers, while advanced composition students compare their scores with others in that class (Smith & Kiefer, 1982). For an example of a student paper and the WWB text-analysis programs, see Appendix B.

The Writer's Workbench and Native-Speaker Composition Students

In the fall of 1981, Smith and Kiefer (1982) began adapting the WWB for collegiate use and began student testing the text-analysis programs. With the help of Blake Stewart, a programmer at the CSU Computer Center, 17 WWB programs were adapted and joined to run with one command. The DICTION program was modified, new programs called DEVELOPMENT and VAGUENESS were tested, and standards for PROSE style were established based on 40 holistically scored superior freshmen essays. At the same time, Kiefer and Smith (1984a) began studying the effects of WWB use on students' editing and writing skills. In the pilot study, 38 students enrolled in two sections of freshman composition spent at least 1 hour a week using the computer to type their essays and to run 4 WWB text-analysis programs (DICTION, SUGGEST, SPELL, and STYLE) on their writing. Freshmen in two control sections of composition also practiced the writing and editing skills, using the same textbook and syllabus, but did not use the Computer Lab. All the students were given a two-part editing pre-test task that required them to (a) identify errors in mechanics and usage and to (b) revise a passage for simplicity and clarity. Students in the experimental section also completed an attitude survey on the use of computers in composition classes.

On the two post-test editing tasks, students from both the experimental and the control classes identified approximately the same number of errors in the first task (skills taught in all classes during the semester); however, on the second task, there were significant differences between the groups in revising for simplicity, directness, and clarity, the editing skills taught by DICTION, SUGGEST, and STYLE (Kiefer & Smith, 1984a, p. 206). On this task, the experimental students outperformed the control sections by 64%. Moreover, student attitudes toward the computers were overwhelmingly positive; 65% responded that they would "look forward to another composition class if it used computers" (Kiefer & Smith, 1984a, p. 206).

Kiefer and Smith (1984a) concluded that textual analysis was a potential teaching tool that could *teach* editing skills. It appears that

the interactive qualities of the WWB motivated students to improve their writing. "Our tests suggest that textual analysis with computers intrigues college writers and speeds learning of editing skills by offering immediate, reliable, and consistent attention to surface problems" (1984a, p. 210).

During the next two years, additional research by Kiefer and Smith (1984b) in an expanded Computer Lab involving greater numbers of freshmen composition students and students enrolled in other courses as well as composition teachers showed similarly positive attitudes toward use of the word processor and the text-analysis programs. Students in College Composition consistently improved 40% to 50% more on tests of editing skill than did their counterparts in control groups, even when the same teacher followed the same syllabus (see Frase, Kiefer, & Smith, in press).

The Writer's Workbench and Holistically Scored Writing Samples

As part of the continuing integration of the WWB into CSU's composition program, Reid and Findlay (1984) began, in 1983, to study the ability of the WWB programs to measure the quality of essays.[3] They selected 44 holistically scored essays from the 1982 administration of the CSU Composition Placement Examination that is given to each entering freshman. Reid and Findlay divided the essays evenly across the range into low (scores 1-3), middle (scores 4-6) and high (scores 7-9) score categories. Each selected essay had previously received three independent scores that varied no more than 1 point; thus they represented a reliable evaluation of overall writing quality that could be compared with the WWB numerical descriptions of the essays.

Statistical analysis determined which WWB stylistic measures, singly or in combination, were most predictive of essay quality. Nine of the 27 WWB measures were statistically significant ($p < .05$). The measure most highly correlated with the holistic rations was Essay Length ($r = .70$). Other individual quantitative measures exhibiting high correlations with the holistic ratings were number of Spelling errors ($r = .57$), Kincaid Readability score ($r = .57$), Average Word Length ($r = .52$), Percentage of Content Words ($r = .39$), and Average Sentence Length ($r = .38$). Other measures showing lower though still significant correlations with holistic ratings included Percentage of Pronouns ($r = .35$). Percentage of Short Sentences ($r = .31$), and Percentage of Abstract Words ($r = .28$).

The results of a multiple regression analysis showed that several combinations of WWB programs are potentially useful in evaluating student essays. The combination of Essay Length, Kincaid Readabil-

ity, and Spelling errors (adjusted R^2 = .84) and the combination of Essay Length, Spelling errors, and Sentence Length (adjusted R^2 = .83) suggest the validity of the WWB for native-speaker essay scoring.

Reid and Findlay reported that some quantifiable WWB measures of STYLE do correlate highly with holistically evaluated student essays. However, they concluded that modifying many of the standards for student writing would make the WWB less conservative and more helpful. For example, they suggested increases in the appropriate Readability range, Sentence Length range, and in the allowable percentages of ABSTRACT words, "To Be" verbs, Nominalizations, VAGUE words, Passive Voice, and Subject Openers (e.g., sentences that begin with nouns, or pronouns).

The Writer's Workbench and English as a Second Language (ESL) Writers

Based on the positive results with native-speaker essays, CSU faculty in the Intensive English Program began to investigate WWB use with ESL students. In the spring of 1982, 18 international students in the Intensive English Program participated in a pilot study using the Computer Lab in conjunction with their writing classes (Reid, Lindstrom, McCaffrey & Larson, 1983). Pre- and post-writing samples, editing samples, and affective questionnaires were administered at the beginning and end of the 7-week writing classes. Results of the project demonstrated that international students were capable of using the word processing program (though some elected to enroll concurrently in an ESL typing class to increase their efficiency on the terminals). Even with only half a semester of exposure to the computers, student writing improved slightly over that of the control group composed of students in the writing classes who were not using the computers. Results of the questionnaires showed that the ESL students felt that their writing skills had improved as a direct result of using the word processor and the WWB.

The success of the study by Kiefer and Smith and the ESL pilot project described above led to the writing of a grant proposal that would assist in continuing the ESL-WWB research. During the fall of 1983, the Exxon Education Foundation funded a study of the impact of the WWB on international student writing. Two noncredit "grant classes" were offered during the spring '84 and fall '84 semesters. Full-time undergraduate and graduate international students, at the university were eligible. Participants spent 1 hour of a 3-hour per week class in the Computer Lab, while the other class hours were devoted to the theory and practice of writing U.S. academic prose. (An on-line

LEARN program, developed for the WWB by Charles Smith, enabled students to develop proficiency with the word processor in less than an hour.)

Data collection was modelled on the Kiefer and Smith native-speaker studies. Pre- and post-writing topics, originally developed for a study by Sybil Carlson and Brent Bridgeman (1983) of Educational Testing Service (ETS) and used in preliminary studies of the ETS TOEFL essay test[4], were administered to grant class participants. Students provided two 30-minute samples representing different styles of writing at the beginning and the end of the semester for a total of four samples. One essay (called Graph for the purpose of analysis) asked students to describe a chart. These two writing samples were based on the interpretation of (a) a bar graph that depicted changes in farm size and population and (b) a pie graph that distinguished between percentage of world population and percentage of land mass on each continent. The other essay task (called *Other*) asked students to discuss a comparison/contrast topic. The topics included (a) two ways of spending leisure time and (b) the advantages and disadvantages of space exploration.

During the semester, the grant class participants wrote weekly, practicing the paragraph and essay form as well as analytic, technical, and research writing. All writing was entered by the students on the computers, and the resulting text-analysis on the printouts was discussed in class and in student-teacher conferences. At the end of the semester, the pre- and post-writing samples (randomized, with student names removed) were scored holistically, using a 1-6 scale. Each paper received two independent readings. Papers whose initial scores differed by more than 1 point were read a third time. The interrater reliability for the holistic scoring was .79.[5]

In a matched *t*-test, the holistic scores for pre- and post-writing samples showed statistically significant improvement for the students who participated fully in the grant classes (Table 1).[6] Students who enrolled in the class, but who missed at least half of the coursework, formed a contrasting group. These students felt that they needed to improve their writing but were unable to commit the necessary time to a noncredit class. Although these students scored higher on the *pre*-tests than the general population in the class, their *post*-tests received lower holistic scores.

While it is not possible to infer from these results that the use of word processor and text-analysis programs was solely responsible for the improvement in student writing skills, it is clear that a class that focused on improving international student writing skills and that used the WWB did result in significantly improved student writing.

In addition to the pre- and post-writing samples, students in the

Table 1
Matched t-Test of Holistic Scores for Pre- and Post-Writing Samples of ESL Students

Participants*		Contrasting Group**	
pre-Graph X = 3.78	post-Graph X = 4.44	pre-Graph X = 4.82	post-Graph X = 4.36
t-value = 3.11 at <.005		t-value = -.9847 n.s.	
pre-Other X = 4.64	post-Other X = 5.14	pre-Other X = 5.0	post-Other X = 4.68
t-value = 3.78 at <.005 +		t-value = -.9697 n.s.	
*N = 25		**N = 11	

grant class completed pre- and post-questionnaires concerning the use of the computer at the beginning and end of the semester, and they completed a post-course affective questionnaire. These student evaluations of the grant class (Table 2) strongly indicate that they recognized the need for additional writing assistance, that they responded positively to the Computer Lab, and that they appreciated the feedback they received from their teacher during the class. In written comments about the class, students unanimously suggested that such a class be offered for credit so that, as two students commented, the class can "provide more time for computer work," and students can "give (the class) the priority it deserves." Faculty advisers of grant class students across the University were given the opportunity to comment on any changes in their students' writing; of the total number that replied (43), 42% of those who had knowledge of their students' writing reported that they had seen improvement in their writing skills, and none reported a negative change.

Finally, more than half (59%) of the students responding to the questionnaire concerning the computer text-analysis programs believed that their writing skills improved because of using the computer, 87% felt that they would be able to use what they learned from using the computer in their writing for other classes, and 93% found using the computer text-analysis an enjoyable experience. Individual comments about the most helpful WWB programs varied widely, indicating that the text-analysis programs respond to diverse needs.

Table 2
ESL Student Evaluation of the Grant Classes

	Not useful	Somewhat useful	Useful	Very useful	Cannot answer	N
computer lab work	0	1 4.35%	12 52.17%	10 43.48%	0	23
computer programs	0	0	11 47.83%	9 39.13%	3 13.04%	23
class discussions	0	0	10 43.48%	12 52.17%	1 4.35%	23
teacher	0	0	2 9.09%	20 90.91%	0	22
textbook	0	1 4.35%	9 39.13%	13 56.52%	0	23
writing assignments	0	0	10 43.48%	13 56.52%	0	23
writing practice	0	0	10 43.48%	13 56.52%	0	23

The Writer's Workbench and Holistically Scored ESL Student Essays

At the end of the grant classes, the pre- and post-writing samples for all grant class enrollees[7] were keyed into the computer and analyzed by type (*Graph* vs. *Other*) in an effort to determine whether the two kinds of writing were quantifiably different. Two Bell Labs WWB programs, MKSTAND and STYLEFILES[8] were run on the directories to determine the means of the various WWB programs for each type's STYLE variables.[9] An independent t-test on the resulting data showed that *Graph* and *Other* writing samples differed significantly ($p < 0.05$) in five of eight WWB measurements (Table 3). In other words, in the 64 *Graph* and 69 *Other* ESL student writing samples, those that described and analyzed graphs differed from those used to explain and discuss contrasting ideas. Generally speaking, ESL students wrote shorter *Graph* essays, but the Average Sentence Length was longer. Average Word Length was slightly shorter in the *Graph* essays, while

the Percent of Content Words was higher. Finally, ESL writers used only half as many pronouns in the *Graph* essays as they did in the *Other* samples. Three other WWB programs – Readability; Percentage of Simple and Complex Sentences; and Percent of To Be and Passive Voice verbs – did not differ significantly between the two types of writing samples. Surprisingly, the holistic scores assigned also showed a significant difference between *Graph* and *Other*; holistic evaluators rated the *Other* essays significantly higher than the *Graph* samples.

Table 3
Independent t-Tests of Five WWB Measures Using Graph and Other Writing Samples[10]

WWB Measurements	Mean	Standard Deviation	t Value	2-Tail Prob.
**Number of words in the essay				
Other	198.2319	67.110	4.58	.000
Graph	151.0469	51.249		
*Average sentence length				
Other	19.7768	5.477	-2.32	.022
Graph	22.1797	6.434		
*Average word length				
Other	4.7399	.305	1.99	.049
Graph	4.6373	.288		
*Percent of content words				
Other	56.8333	3.975	-3.49	.001
Graph	59.1578	3.690		
*Percent of pronoun use				
Other	6.8043	2.908	6.56	.000
Graph	3.8016	2.308		
*Holistic Evaluations				
Other	4.5870	1.154	2.98	.003
Graph	4.0156	1.046		

*Pooled variance estimates: df = 131 Other N = 69
**Separate variance estimate: df = 126.44 Graph N = 64

The *Graph* and *Other* writing samples were then subdivided into high (score: 5-6), middle (score: 3-4), and low (score: 1-2) score categories to determine whether certain WWB programs could distinguish holistically determined quality writing from middle range and low holistic scores. Correlational analysis revealed that several WWB programs were predictive of essay quality (Table 4). In both the *Graph* and the *Other* testing samples, Essay Length, Average Word Length, and Kincaid Readability correlated significantly with quality essays. Other WWB measurements, such as Number of Content Words, Percentage of Compound Sentences, and Percentage of Short Sentences, showed lower but still significant correlations in one or the other testing sample.

Table 4
Correlation of WWB Factors to Holistically Scored ESL Essays

WWB Measurements	Holistic Score Other
Essay Length	r .5188 p <.001
Average Word Length	r .3396 p <.002
Kincaid Readability	r .2749 p <.011
Percent of Complex Sentences	r -.1675 p <.084
Percent of Content Words	r .1597 p <.095
	Holistic Score Graph
Essay Length	r .5715 p <.001
Average Word Length	r .4528 p <.001
Percent of Pronoun Use	r -.2149 p <.044
Percent of Short Sentences	r .2022 p <.055
Kincaid Readability	r .2071 p <.050

Stepwise multiple regression revealed that combinations of the positively correlated WWB factors could validly predict quality essays (Table 5). Specifically, for the *Graph* and the *Other* samples, Essay Length and Average Word Length accounted for 41% and 35% respectively of the variability in holistically scored student essays. That variance of other significantly correlated measures in the single correlation studies (Readability; Percent of Short, Long, and Complex Sentences; and Percent of Content Words) overlapped with the variance for the two most highly correlated factors.

It is interesting to note that some ESL-WWB research results parallel previous WWB research with native-speaker compositions. In the Reid and Findlay analysis, Essay Length, Average Word Length and Kincaid Readability also correlated with essay quality. Moreover, in research that investigated essays by experienced writers in popular freshman reading-writing textbooks (Garvey & Lindstrom, 1984), the professional samples differed significantly from freshman compositions in Kincaid Readability, Average Word Length, and Average Sentence Length. Surprisingly, in both the native speaker and the ESL research, several of the more sophisticated WWB measures, such as a high Percentage of Compound-Complex Sentences or a low Percentage of Passive Voice, did not correlate with quality writing.

Table 5
Multiple Regression Analysis of WWB Factors and Holistically Scored Essays

Graph	
Essay Length	(adjusted R^2
Average Word Length	= .41158)
Other	
Essay Length	(adjusted R^2
Average Word Length	= .35067)

Conclusion

The findings obtained from WWB research at CSU warrant the following conclusions.

1. Both native-speaker freshman composition students and ESL composition students who used the WWB as an integral part of their writing classes improved their editing and writing skills significantly.

whereas students who did not have access to the WWB did not improve their skills.

2. Several WWB programs, singly or in combination, are potentially useful in evaluating both native and non-native student essays; the most predictive programs are Essay Length, Kincaid Readability scores, and Average Sentence Length.

3. For ESL student writing, the WWB quantitative analysis shows that the two types of writing studied—describing/interpreting graphs and explaining/discussing contrasting ideas—differ with respect to several text characteristics.

4. Both native speakers and ESL students, in a variety of writing classes, as well as teachers of those classes, report positive attitudes towards the WWB word processor and text-analysis programs. The diversity of comments concerning "the most helpful" WWB programs indicate that the programs respond to a wide range of student needs.

In general, student use of the WWB at CSU has increased each year and has been particularly successful because of the interactive quality of the software programs, the individualized printout information, and the programming capability for different sets of standards for different types of writing and levels of writing ability. At present, 4,000 composition students use the Computer Laboratory and the WWB each school year. The Computer Lab has expanded, first to four Onyx computers with seven terminals each and more recently to include an AT&T 3B2 computer supporting five terminals. Future plans include a total of 44-60 terminals and the acquisition of newly developed pre-writing software so that the use of the Computer Lab will be available to all CSU students.

Further research in many pedagogical and evaluative areas remains to be done. For example, the links between the cognitive development of writing skills and the usefulness of the WWB in that development need investigation. Are some WWB programs more helpful for remedial composition students, for experienced writers, for ESL students? Research and adaptation of WWB programs is already underway by the CSU Department of Business for business communication classes (Johnson & Sterkel, 1983). Other fertile areas of research include longitudinal studies concerning the lasting effects of the WWB on student writing, and studies of the validity and reliability of holistic evaluations and WWB programs. WWB quantitative information identifying differences in the presentation of written material among differing language and cultural backgrounds could provide a founda-

tion for further discourse analyses and contrastive rhetoric studies. Finally, research should be conducted using the WWB to investigate the quantifiable differences in various types of writing—for example, narrative, exposition, argumentation, scientific and technical prose. Such research would provide baseline data for computer-adaptive teaching and testing and for the setting of valid standards for the teaching and testing of composition.

Appendix A
Description of WWB Programs Used at CSU

ORGANIZATION prints the first and last sentence of each paragraph in the student's text to give the writer an outline of the essay. Students check the output for unity and coherence.

TOPIC prints the most frequently used content words with the number of times each is used in a paragraph or essay. ESL students are the primary users of this program; they check the output for focus.

DEVELOPMENT counts words in each paragraph and compares the results with averages drawn from sample papers from the course. If the figures are significantly lower than average, the program suggests that the student check for adequate detail.

FINDBE capitalizes and underlines all forms of *to be* in the text. Native-speakers are asked to limit the percentage of *to be* verbs (to replace some with more active verbs); ESL students are asked to check the subject-verb agreement for each capitalized to be verb.

DICTION flags (i.e., capitalizes and encloses in brackets) any of about 500 wordy, overused, and misused words and phrases in the student text.

SUGGEST immediately follows the text with advice concerning the DICTION flags. Brief SUGGESTions are given on the printout; longer explanations are indicated by referring the student to the *Glossary*. Students must then check the last half of their computer manuals for the longer suggestions.

SPELL lists typographical and spelling errors (that is, words not found in the WWB dictionary). Students must check a dictionary to correct misspelled words.

VAGUENESS counts any of 170 vague or general words. If the text has more than 5% of these words, the printout lists them and recommends revisions. If the text has fewer than 5%, students see only the percentage.

CHECK lists commonly confused homophones and word pairs when the writer has used one of the words in the text. The program includes a brief distinction between the words or a reference to a longer explanation in the *Glossary*.

PUNCTUATION checks for missing parentheses and for patterns of punctuation: periods followed by capital letters, commas and periods inside quotation marks, etc. It lists both the abnormal sentence as punctuated in the text and a suggested change.

GRAMMAR identifies most split infinitives and misuses of *a* and *an*; it lists the sentence with the error and a suggested change.

PROSE compares values for 10 stylistic criteria in a student's paper with standards derived from excellent papers in the course. The criteria include average sentence length, percentages of short and long sentences, sentence types, and percentages of verb types and sentence beginners. When a student's figures fall outside a range of +/- one standard deviation from the mean, the program makes suggestions for improvement.

PASSIVE prints out all passive sentences appearing in a student's text with a reminder to use only effective passive constructions and to change ineffective passives to active voice.

NOMINALIZATION prints out all sentences with nominalized words (i.e., words ending in -ance, -ence, -ion, or -ment) when the percentage of nominalization exceeds 3% of the total number of words. Students are advised to use specific detail and to check for too many nominalizations.

STYLE summarizes numerical information about sentence length and type, as well as about sentence openings and word class counts; the summary includes the total number of words and sentences, and four readability scores: Kincaid, Coleman-Liau, auto, and Flesch.

ABSTRACT compares words in a text to a dictionary of 314 abstract words (determined through psycholinguistic research). If students use more than 2.3% of these words in a text, the program prints a message reminding students to check for adequate concrete detail and lists abstract words in the text.

Appendix B
Writer's Workbench Student Sample

Organization

One thing I feel that irritate me about American youth is that every time I go to a movie, they are shouting, laughing, and talking loudly during a movie. Before I came to America, people said that Americans were polite, and cared about public morals, but now I doubt that and really hate being disturbed by American youths during movies.

FINDBE
(Check all Capitalized *to be* Verbs For Subject-Verb Agreement)

One thing I feel that irritate me about American youths *IS* that every time I go to a movie, they *ARE* shouting, laughing, and talking loudly during a movie. When I first arrived in the U.S. and went to a movie, I thought that *[MAYBE]* only a few American youths had this habit, but after several times, it happened again and again, so I began to believe that most of American youths have this bad habit, both men and women. For example, when they feel the expression of an actor *IS* not suitable for the plot, or when the drama makes no sense, or especially when the movie concerns sex, they shout some words, laugh loudly, and even whistle loudly. At the theater, their actions really disturb others in the audiences, but these American youths act naturally and on purpose. Before I came to America, people said that American WERE polite, and cared about public morals, but now I doubt that and really hate being disturbed by American youths during movies.

SUGGEST

maybe: GLOSSARY

SPELLING

No spelling errors found in tsengd

TOPIC

4 American youths
3 movie
1 women
1 this habit
1 these American
1 theater
1 talking
1 some words
1 several times
1 purpose
1 public
1 plot
1 people
1 one thing
1 no sense
1 movies
1 movie concerns
1 laughing
1 laugh
1 irritate

VAGUENESS

In this text, 6.3 percent of the words are vague.

Below is a list of the vague words in your text. Delete where possible. Otherwise, replace or supplement with specific words, phrases or examples so that your index is as low as possible, preferably below 3%.

thing	go	few	several	most
bad	some	really	people	really

CHECK

bad: GLOSSARY

PUNCTUATION

A description of the punctuation in tsengd:

0 double quotes and 0 single quotes
0 apostrophes
0 left parentheses and 0 right ones

Sentences probably incorrectly punctuated followed by corrections:

No errors found in tsengd

GRAMMAR

For file tsengd:

No split infinitives found

PROSE

VARIATION

Variation in sentence length, type, and openings prevents monotony. More important, a lack of variation suggests that every topic and every sentence has equal weight – a deficiency making it difficult for the reader to pick out the important points.

In this text 20% of the sentences are simple, 80% are complex, giving a difference of -60. This difference should range from -40 to 10 for good papers of this type.

This paper contains many more complex sentences than is common for papers in C0150. (Complex sentences contain an independent and a dependent clause.) One way to improve this text is to rephrase the most important ideas in simple sentences. Doing so will give those ideas additional emphasis. For help, see page ____ in the Manual.

Additionally, the longest sentence is 51 words long. Each of the sentences below is over fifty words long. Sentences this long may be lists. If so, consider a list format to make them easy to follow. If they are not lists, check for clarity: you may want to break some or all into shorter segments.

When I first arrived in the U.S. and went to a movie, I thought that maybe only a few American youths had this habit, but after several times, it happened again and again, so I began to believe that most of American youths have this bad habit, both men and women.

Especially if this is a scientific or technical text, such changes would also help to reduce the average sentence length. Your average is 34 words. A good average would be 16 to 26 words.

SENTENCE STRUCTURE

Passives

Your text contains 4.0% passive voice verbs.

Good papers in this course typically contain fewer then 5%. Check each sentence below for effective use of the passive, and convert all ineffective passives to active. For help, see your manual, PROSE, Passive Voice.

before I came to America, people said that Americans were polite, and cared about public morals, but now I doubt that and really hate being disturbed by American youths during movies.

Nominalizations

You have appropriately limited your nominalizations (nouns made from verbs, e.g., "description").

STYLE

sentence info:
av sent leng: 34.4
% of sent 5 words shorter than av: 20% (1)
% of sent 10 words longer than av: 20% (1)
sentence types
simple 20% (1) complex 80% (4)
verb choice
to be as percent of total: 21% (5) aux 0% (0) inf 4% (1)
passives as % of non-inf verbs 4% (1)
nominalizations 1% (2)
sentence beginnings:
subject openers: noun (0) pron (0) pos (0) adj (1) art (0) TOTAL 20%
other openers: prep 40% (2) adv 0% (0)
verb 0% (0) sub conj 40% (2) conj 0% 0
other information
no. sent: 5; no. wds: 172
av word leng: 4.41
no. questions: 0 no. imperatives: 0
no. content wds: 98 57.0% av leng: 5.51
word types as % of total
prep 8.7% (15) conj 7.6% (13) adv 8.1% (14)
noun 23.3% (40) adj 13.4% (23) pron 8.1% (14)
readability
(Kincaid) 16.2 (auto) 16.6 (Coleman-Liau) 9.3 (Flesch) 14.5 (40.0)

ABSTRACT

In this text, 5.0 percent of the words are abstract.

Psychological research shows that concrete texts are easier to read, easier to use, and easier to remember. Your abstract index—over 2%—suggests that your text might be strengthened by adding more specific details and concrete illustrations. The following list of abstract words — in order of their appearance — may help you identify sections were support is necessary.

Your abstract words are:

american	thought	american	believe	american
american	public	american		

Run on Thu Mar 21 09:29:38 MST 1985.
File = tsengd Run by user htsengd
English Department ONYX Number Four

References

Carlson, S., & Bridgeman, F. (1983). *Survey of academic writing tasks required for graduate and undergraduate foreign students*. TOEFL Research Report (No. 15). Princeton, NJ: Educational Testing Service.

Cherry, L. (1982). Writing tools. *IEEE Transactions of Communication, 30*(1), 100-105.

Cherry, L., Fox, M. L., Frase, L. T., Keenan, S. A., & MacDonald, N. H. (1983). Computer aids for text analysis. *Bell Laboratories Record*, May/June, 10-16.

Frase, L. T., Kiefer, K. E., & Smith C. R. (in press). Theory and practice in computer-aided instruction. In R. Freedman (Ed.), *Evaluation, response and revision*. Norwood, NJ: Ablex.

Garvey, J., & Lindstrom, D. (1984). Pro's prose and the Writer's Workbench. Paper presented at the annual College Composition and Communications Conference, New York.

Johnson, M., & Sterkel, K. (1983). Writing with a computer brings positive results. *Business Education Forum, 38*(5), 32-36.

Kiefer, K., & Smith, C. (1984a). Improving students' revising and editing: the Writer's Workbench system. In William Wresch (Ed.), *The computer in composition instruction: a writer's tool* (pp. 65-82). Urbana, IL: National Council of Teachers of English.

Kiefer, K., & Smith, C. (1984b). Textual analysis with computers: tests of Bell Laboratories' computer software. *Research in the Teaching of English, 17*(3), 201-214.

MacDonald, N. H. (1980). Pattern, matching, and language analysis as editing supports. In L. T. Frase (Chair), *Computer aids for text editing and design*. Symposium conducted at the annual meeting of the American Educational Research Association, Boston.

MacDonald, N. H., Frase, L. T., Gingrich, P., & Keenan, S. (1982). The Writer's Workbench: computer aids for text analysis. *IEEE Transactions on Communications, 30*(1), 105-110.

Reid, J., Lindstrom, M., McCaffrey, M., & Larson, D. (1983). Computer-assisted text for ESL students. *CALICO Journal, 1*(3), 40-46.

Reid, S., & Findlay, G. (1984). Writer's Workbench analysis of holistically scored essays. Manuscript submitted for publication.

Smith, C., & Kiefer, K. (1982). Writer's Workbench: computers and writing instruction. *Proceedings of the Future of Literacy Conference*. University of Maryland, Baltimore: Baltimore County.

Footnotes

[1]My thanks to the Exxon Education Foundation for the opportunity to investigate international student use of the WWB.

[2]Within this article, the titles of individual WWB programs are fully capitalized, separate parts of individual programs have their first letter capitalized.

[3]The information and statistical data in this section are used with the permission of the researchers.

[4]Because the CSU Department of English was also involved during 1983 in typing 400 sample TOEFL essays into the computer and in running the WWB programs on those essays, Sybil Carlson, director of the TOEFL-funded writing research project, gave permission for the grant classes to use the TOEFL topics. I am grateful to her for expanding the base of my data.

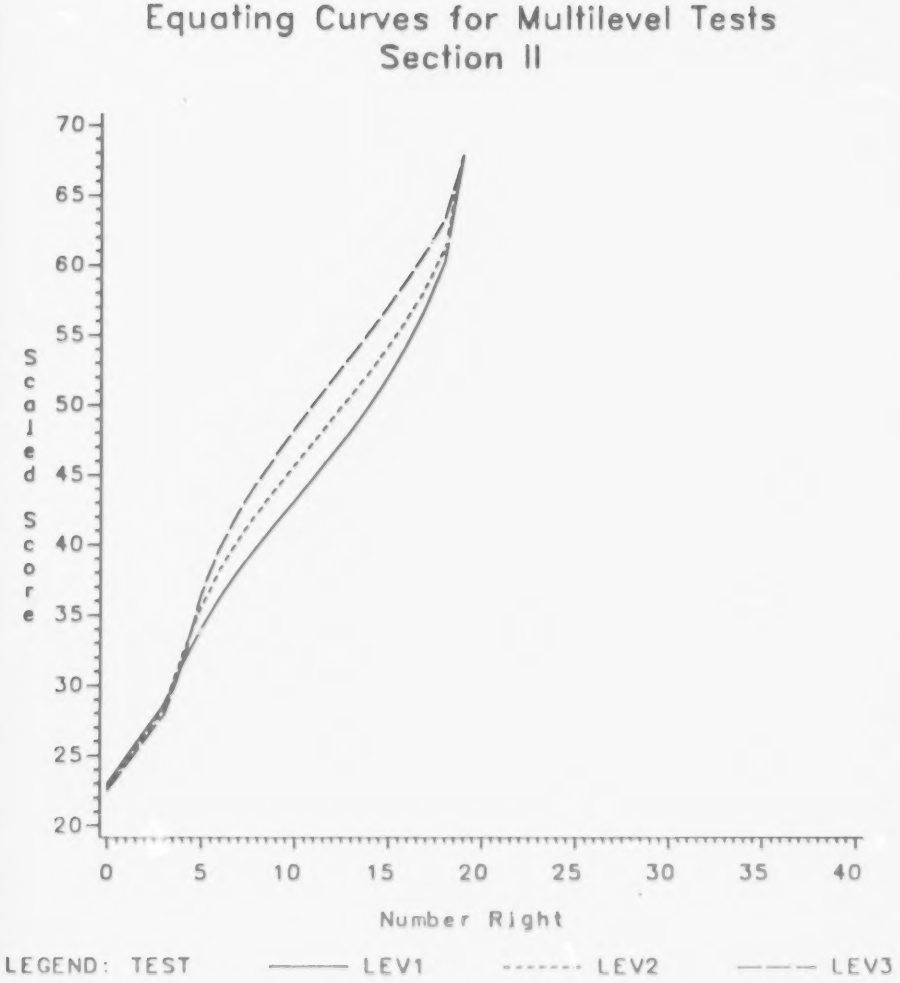
[5]Peggy Lindstrom, Curriculum Supervisor in the Intensive English Program, and Eloise Ariza, MA TEFL candidate at CSU, assisted in teaching the grant classes and in the holistic evaluation of the pre- and post-writing samples.

[6]For these tests of significance, only the writing samples for grant class students who had completed both the pre- and the post-test samples were used.

[7]The total number of holistically scored essays in this statistical test was greater than in the original t-test because some students took the pre-test but not the post-test.

[8]MKSTAND collates some STYLE scores from selected essays and gives ranges and standard deviations for each score. STYLE-FILES gives individual STYLE scores for each paper.

[9]The STYLE program gives the following information: four readability scores, number and types of sentences (questions, imperatives), number of content words, average sentence length, percentage of short and long sentences, percentage of sentence types (simple, complex, compound, compound-complex), percentage of word types (verbs: to be, auxiliary, infinitive, and passives; prepositions, conjunctions, adverbs, nouns, adjectives, pronouns, and nominalizations), and percentage of sentence beginnings (subject openers: nouns, pronouns, possessives, adjectives, and articles; prepositions, adverbs, verbs, subordinate conjunctions, conjunctions, and expletives).

[10]Three other comparisons were performed and were found to be nonsignificant at the .05 level. Note that the majority of the comparisons were significant, many of those at the .01 level.

Erratum: Page 86, Figure 1. For Equating curves for multilevel tests: Section III
Read Equating curves for multilevel tests: Section II
(figure below)